SESSOMS, JOHN CAMERON LEE, Ph.D. Level-specific Fit Index Performance with Diagonally Weighted Least Squares Estimation of Multilevel Structural Equation Models. (2019)
Directed by Dr. John T. Willse. 239 pp.

Level-specific fit indices generally are advised (but rarely used in practice) for evaluating fit of multilevel structural equation models (MSEM). Level-specific fit indices assess model fit for each level separately. Aggregate fit indices combine (mis)fit across levels and are used predominantly in practice. Diagonally Weighted Least Squares (DWLS) estimation generally is recommended for MSEM with categorical variables. Previous evaluations of level-specific fit indices only used continuous, normally distributed variables, small models, and Maximum Likelihood estimation. However, MSEM applications often use large models and categorical, non-normal variables. Single-level DWLS fit indices usually become less sensitive to misfit as model size and distributional skew/asymmetry increases.

This simulation study evaluated how categorical variables, distributional skew, Level-2 sample size, intraclass correlation, model size, and model misspecification affected level-specific and aggregate fit indices. Level-1 and aggregate fit indices usually performed similarly. Level-1 and aggregate fit indices usually detected large Level-1 misfit but not small Level-1 misfit. Aggregate fit indices never identified Level-2 misfit. Level-1 and aggregate fit indices never rejected correct Level-1 models. Level-2 fit indices usually had low power to reject Level-2 misfit except with very optimal data. Level-2 fit indices often rejected correct Level-2 models. Researchers likely should consider alternatives to Level-2 fit indices and MSEMs.

LEVEL-SPECIFIC FIT INDEX PERFORMANCE WITH DIAGONALLY WEIGHTED LEAST SQUARES ESTIMATION OF MULTILEVEL STRUCTURAL EQUATION MODELS

by

John Cameron Lee Sessoms

A Dissertation Submitted to
the Faculty of The Graduate School at
The University of North Carolina at Greensboro
in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Greensboro
2019

Approved by

Dr. John Willse
Committee Chair

DEDICATION

My dissertation is dedicated to my mom Judy Sessoms.

APPROVAL PAGE

This dissertation written by JOHN CAMERON LEE SESSOMS has been approved by the following committee of the Faculty of The Graduate School at The University of North Carolina at Greensboro.

Committee Chair _____
Dr. John Willse

Committee Members _____
Dr. Robert Henson

_____
Dr. Ric Luecht

_____
Dr. Kyung Yong Kim

_____
Date of Acceptance by Committee

_____
Date of Final Oral Examination

ACKNOWLEDGEMENTS

Many people have supported me during my life, and I wouldn't have accomplished my PhD without them.

First, I want to thank my dissertation chair and advisor Dr. John Willse. Thank you for your feedback and guidance. I always enjoy our interactions. You always asked thought-provoking questions to which I often did not have an immediate answer. You helped me think much more deeply about quantitative methods.

Second, I want to thank my dissertation committee: Dr. Kyung Yong Kim, Dr. Bob Henson, and Dr. Ric Luecht. Kyung Yong, thank you for your very helpful dissertation feedback. I haven't had the privilege of taking any of your classes, but I was very impressed with your presentation and knowledge when you interviewed at UNCG. Bob, you are one of the most brilliant people I've ever met. Your knowledge of countless equations and ability to write them up at a moment's notice is astounding. Thank you for being very supportive and understanding. I appreciate your willingness to meet with me whenever about anything. Your advice and discussion helped me overcome several huge obstacles with the dissertation. Ric, I've always been amazed at how much you know and your many significant contributions to educational measurement. We always had a blast during class. Thank you for all that you've taught me and the good-natured ribbing during class that always made me laugh. You helped me better understand the "big picture" and difficulties of using statistical methods in practice. I will always remember "it depends" when I make decisions in my career.

I need to thank my Masters' advisor Dr. Sara Finney at James Madison University. I never thought I would meet someone who loves structural equation models more than I do… until I met you. Thank you for all of your excellent feedback that greatly improved my writing, critical thinking, and relationships. Every day you inspire me to study more, learn more, and constantly assess my statistics knowledge. No one has had a bigger impact on my professional development than you.

I also need to thank some of my undergraduate professors at Liberty University. Dr. Fred Volk, thank you for pushing me to work hard and take research seriously. Thank you for always meeting with me to provide advice, even if the meeting was unscheduled. Thank you for introducing me to factor analysis, which was the catalyst for me becoming interested in statistics. Factor analysis remains tied with structural equation models as my all-time favorite quantitative method. Dr. Brianne Friberg, thank you for introducing me to structural equation models. Your meeting with me to discuss their value and letting me run path analysis on Mplus has made all the difference in my career. Without you, I wouldn't have been able to get in to JMU. Thank you for all the research advice and personal guidance you gave during my time at Liberty. Thank you for always making time for me and always supporting me.

Finally, I need to thank my parents and brother. Peyton, thank you for taking me on a college visit to Liberty. Dad, thank you for all of the support during college and grad school. Mom, thank you for all you do for me. Your constant love and support have made all the difference.

TABLE OF CONTENTS

Page

LIST OF TABLES ix

LIST OF FIGURES x

CHAPTER

I. INTRODUCTION 1

Prevalence of Attitudinal Measures 1
Difficulties of Modeling Attitudinal Measures 1
Structural Equation Models 2
Multilevel Structural Equation Models 5
Need for the Current Study 6

II. LITERATURE REVIEW 8

Structural Equation Models (SEM) 8
Multilevel Structural Equation Models (MSEM) 10
MSEM Literature Overview 16
Model Fit Evaluation in MSEM: Importance of Level-Specific Fit 21
Yuan and Bentler's Approach to Level-Specific Fit 22
Ryu and West's Approach to Level-Specific Fit 23
Aggregate vs. Level-Specific Fit in Application 29
Level-Specific Fit Indices' Performance 29
MSEM Estimation Methods 30
Mathematical Definition of WLS and DWLS Estimation 33
DWLS Performance 35
MSEM Research Limitations on DWLS Performance 38
Literature Review Summary and Need for Current Study 42
Research Questions 43

III. METHODS 45

Overview of Methods 45
Simulation Conditions 45
Evaluation of Results and Analytic Approach 71

IV. RESULTS 73

Model Convergence Results 73
Model Fit Index Performance 73
Level-1 Fit Indices' Overall Performance 74
Data Asymmetry's Impact on Level-1 Fit Indices 78
Model Size's Impact on Level-1 Fit Indices 86
Number of Groups' Impact on Level-1 Fit Indices 93
ICCs' Impact on Level-1 Fit Indices 99
Level-2 Fit Indices' Overall Performance 104
Data Asymmetry's Impact on Level-2 Fit Indices 109
Model Size's Impact on Level-2 Fit Indices 117
Impact of Number of Groups on Level-2 Fit Indices 125
ICCs' Impact on Level-2 Fit Indices 134
Aggregate Fit Indices' Overall Performance 143
Data Asymmetry's Impact on Aggregate Fit Indices 149
Model Size's Impact on Aggregate Fit Indices 156
Impact of Number of Groups on Aggregate Fit Indices 161
ICCs' Impact on Aggregate Fit Indices 168
Comparing Level-1, Level-2, and Aggregate Fit Performance 175
Summary of Analysis Results 190

V. DISCUSSION 193

Study Context and Design 193
Defining Fit Index Performance 196
Overall Results Summary 196
Level-1 Fit Indices' Performance 197
Level-2 Fit Indices' Performance 198
Aggregate Fit Indices' Performance 199
Comparing Level-1, Level-2, and Aggregate Fit Performance 201
Comparing Study Results to the Literature 201
Implications for Practice 203
Alternatives to Fit Indices and MSEM 205
Study Limitations and Future Research 206

REFERENCES 208

APPENDIX A. LARGE MODEL CONVERGENCE FOR SKEWED DATA 217

APPENDIX B. LARGE MODEL CONVERGENCE FOR SYMMETRIC DATA 219

APPENDIX C. SMALL MODEL CONVERGENCE FOR SKEWED DATA 221

APPENDIX D. SMALL MODEL CONVERGENCE FOR SYMMETRIC DATA 223

APPENDIX E. EMPIRICAL ASYMMETRY CATEGORY PROPORTIONS 225

LIST OF TABLES

Page

Table 1. Summary of Simulation Conditions for Dissertation.....46

Table 2. Population Parameter Values for Data Generation.....49

Table 3. Thresholds Used to Generate Data Asymmetry Conditions 55

Table 4. Rejection Rates of Level-1 Fit Indices for Small Model Conditions 75

Table 5. Rejection Rates of Level-1 Fit Indices for Large Model Conditions 76

Table 6. Rejection Rates of Level-2 Fit Indices for Small Model Conditions 105

Table 7. Rejection Rates of Level-2 Fit Indices for Large Model Conditions 106

Table 8. Rejection Rates of Aggregate Fit Indices for Small Model Conditions 147

Table 9. Rejection Rates of Aggregate Fit Indices for Large Model Conditions 148

LIST OF FIGURES

Page

Figure 1. Population Small Model that Models 2 Factors at Each Level 50

Figure 2. Population Large Model that Models 4 Factors at Each Level 51

Figure 3. Misspecification Condition 1 for the Small Model 58

Figure 4. Misspecification Condition 1 for the Large Model 59

Figure 5. Misspecification Condition 2 for the Small Model 60

Figure 6. Misspecification Condition 2 for the Large Model 61

Figure 7. Misspecification Condition 3 for the Small Model 62

Figure 8. Misspecification Condition 3 for the Large Model 63

Figure 9. Misspecification Condition 4 for the Small Model 64

Figure 10. Misspecification Condition 4 for the Large Model 65

Figure 11. Misspecification Condition 5 for the Small Model 66

Figure 12. Misspecification Condition 5 for the Large Model 67

Figure 13. Misspecification Condition 6 for the Small Model 68

Figure 14. Misspecification Condition 6 for the Large Model 69

Figure 15. Impact of Data Asymmetry on Level-1 Fit Indices' Rejection Rates with Small Models, Many Groups, and Large ICCs. 79

Figure 16. Impact of Data Asymmetry on Level-1 Fit Indices' Rejection Rates with Large Models, Few Groups, and Small ICCs. 80

Figure 17. Impact of Data Asymmetry on Level-1 Fit Indices' Rejection Rates with Small Models, Many Groups, and Small ICCs. 81

Figure 18. Impact of Data Asymmetry on Level-1 Fit Indices' Rejection Rates with Large Models, Few Groups, and Large ICCs. 82

Figure 19. Impact of Model Size on Level-1 Fit Indices' Rejection of Symmetric Data, Many Groups, and Small ICCs. 87

Figure 20. Impact of Model Size on Level-1 Fit Indices' Rejection of Asymmetric Data, Few Groups, and Large ICCs. 88

Figure 21. Impact of Model Size on Level-1 Fit Indices' Rejection of Symmetric Data, Few Groups, and Small ICCs. 89

Figure 22. Impact of Model Size on Level-1 Fit Indices' Rejection of Asymmetric Data, Many Groups, and Large ICCs. 90

Figure 23. Impact of Number of Groups on Level-1 Fit Indices' Rejection of Large Models, Symmetric Data, and Large ICCs. 94

Figure 24. Impact of Number of Groups on Level-1 Fit Indices' Rejection of Large Models, Symmetric Data, and Small ICCs. 95

Figure 25. Impact of Number of Groups on Level-1 Fit Indices' Rejection of Small Models, Symmetric Data, and Small ICCs. 96

Figure 26. Impact of ICCs on Level-1 Fit Indices' Rejection Rates with Large Models, Symmetric Data, and Few Groups. 100

Figure 27. Impact of ICCs on Level-1 Fit Indices' Rejection Rates with Small Models, Symmetric Data, and Few Groups. 101

Figure 28. Impact of Data Asymmetry on Level-2 Fit Indices' Rejection Rates with Small Models, Many Groups, and Large ICCs 110

Figure 29. Impact of Data Asymmetry on Level-2 Fit Indices' Rejection Rates with Large Models, Many Groups, and Small ICCs 111

Figure 30. Impact of Model Size on Level-2 Fit Indices' Rejection Rates with Symmetric Data, Many Groups, and Large ICCs. 118

Figure 31. Impact of Model Size on Level-2 Fit Indices' Rejection Rates with Symmetric Data, Few Groups, and Large ICCs. 119

Figure 32. Impact of Model Size on Level-2 Fit Indices' Rejection Rates with Asymmetric Data, Few Groups, and Small ICCs. 120

Figure 33. Impact of Number of Groups on Level-2 Fit Indices' Rejection of Large Models, Asymmetric Data, and Small ICCs. 126

Figure 34. Impact of Number of Groups on Level-2 Fit Indices' Rejection of Small Models, Symmetric Data, and Large ICCs. 127

Figure 35. Impact of Number of Groups on Level-2 Fit Indices' Rejection of Large Models, Asymmetric Data, and Large ICCs. 128

Figure 36. Impact of ICCs on Level-2 Fit Indices' Rejection Rates with Small Models, Asymmetric Data, and Few Groups. 135

Figure 37. Impact of ICCs on Level-2 Fit Indices' Rejection Rates with Small Models, Symmetric Data, and Many Groups. 136

Figure 38. Impact of ICCs on Level-2 Fit Indices' Rejection Rates with Large Models, Symmetric Data, and Few Groups. 137

Figure 39. Impact of Data Asymmetry on Aggregate Fit Indices' Rejection of Cross-Loadings Fixed to Zero with Small Models, Many Groups, and Large ICCs. 150

Figure 40. Impact of Data Asymmetry on Aggregate Fit Indices' Rejection of Collapsed Factors with Small Models, Many Groups, and Large ICCs. 151

Figure 41. Impact of Model Size on Aggregate Fit Indices' Rejection of Cross-Loadings with Symmetric Data, Many Groups, and Large ICCs. 157

Figure 42. Impact of Model Size on Aggregate Fit Indices' Rejection of Collapsed Factors with Asymmetric Data, Many Groups, and Small ICCs. 158

Figure 43. Impact of Number of Groups on Aggregate Fit Indices' Rejection of Cross-Loadings Fixed to 0 with Small Models, Symmetric Data, and Large ICCs. 162

Figure 44. Impact of Number of Groups on Aggregate Fit Indices' Rejection of Collapsed Factors with Small Models, Asymmetry, and Small ICCs. 163

Figure 45. ICC Impact on Aggregate Fit Indices' Rejection of Cross-Loadings Fixed to Zero with Small Models, Symmetry, and Few Groups. 169

Figure 46. ICC Impact on Aggregate Fit Indices' Rejection of Cross-Loadings Fixed to Zero with Small Models, Symmetric Data, and Many Groups. 170

Figure 47. ICC Impact on Aggregate Fit Indices' Rejection of Collapsed Factors with Small Models, Asymmetric Data, and Many Groups. 171

Figure 48. Model Size Impact on Level-Specific Fit Index Performance for Level-1 and Level-2 Cross-Loadings Fixed to 0 with Symmetric Data, Many Groups, and Large ICCs. 177

Figure 49. Level-Specific and Aggregate Fit Index Rejection of Correct Model at Level-1 and/or Level-2 with Small Models, Symmetric Data, Many Groups, and Small ICCs. 178

Figure 50. Level-Specific Fit Index Rejection of Collapsed Factors with Small Models, Asymmetric Data, Few Groups, and Small ICCs. 179

Figure 51. Level-Specific and Aggregate Fit Index Rejection of Level-1 Cross-Loadings Fixed to 0 with Small Models, Symmetric Data, Many Groups, and Small ICCs. 180

Figure 52. Level-Specific and Aggregate Fit Index Rejection of Level-1 Collapsed Factors with Small Models, Symmetric Data, Many Groups, and Small ICCs. 181

Figure 53. Level-Specific and Aggregate Fit Index Rejection of Level-2 Cross-Loadings Fixed to 0 with Small Models, Symmetric Data, Many Groups, and Large ICCs. 182

Figure 54. Level-Specific and Aggregate Fit Index Rejection of Level-2 Collapsed Factors with Small Models, Symmetric Data, Few Groups, and Small ICCs. 183

CHAPTER I

INTRODUCTION

Prevalence of Attitudinal Measures

Attitudinal measures are common in psychology and education. These questionnaires and surveys ask respondents their attitudes, opinions, and feelings. These self-report measures are used in a variety of settings for many different topics. High school students may be asked to evaluate their teacher's effectiveness (Sessoms & Willse, 2019). Team members may indicate their team satisfaction (Preacher, Zyphur, & Zhang, 2010). College students may describe how entitled they feel to good grades regardless of academic performance (Sessoms, Finney, & Kopp, 2016). Attitudinal measures have wide applicability and utility.

Difficulties of Modeling Attitudinal Measures

Attitudinal items present scoring and modeling challenges. These measures usually do not have a correct answer. Instead, respondents often are asked to provide the answer that most accurately describes them. Common Likert-style response formats include Disagree/Agree and Not Like Me/Like Me. Affective items often employ 4 to 7 response categories (Li, 2016). Consider an item that uses a Disagree/Agree response format and 4 response option categories. The item states "I enjoy listening to boy bands like One Direction" and measures the degree to which respondents agree. The response categories could be *Strongly Disagree*, *Disagree*, *Agree*, and *Strongly Agree*. Each

response category is considered a valid response. The lack of one correct response means that these items are scored polytomously. A set of affective items often are multidimensional. Structural equation modeling (SEM) typically is used to score and model affective items.

Attitudinal items often are measured in multilevel or nested contexts. Consider students evaluating their teacher's effectiveness. Students are nested within teachers; students with the same teacher likely will provide similar evaluations. Nested data are potentially problematic for researchers. Many statistical analyses such as SEM assume non-nested data. These methods assume that responses are uncorrelated and essentially interchangeable (i.e., independence of observations). In nested data structures, responses from Level-1 units (e.g., students) nested within the same Level-2 unit (e.g., teacher) may be more correlated with each other than with students from different schools. Students with the same teacher tend to be similar to each other and more like each other than students with a different teacher (Stapleton, 2013). Students evaluating the same teacher likely will provide similar evaluations; their evaluations likely will differ from students with a different teacher. Thus, analyzing nested data with methods that assume non-nested data can result in underestimated standard errors and biased parameter estimates (Peugh, 2010).

Structural Equation Models

SEM is a family of statistical models. The three general SEMs are path models, confirmatory factor analysis (CFA) models, and structural regression models (Kline, 2011). Path models are used to estimate causal effects (e.g. direct and indirect effects)

between observed variables. Path models are appropriate when assessing the observed relationship between constructs that are each measured using one observed variable (e.g. one item). Path models could be used to estimate whether extraversion directly affects enjoyment of college. Both variables are measured using one item (e.g., "I am an extrovert" and "I enjoy college").

CFA is appropriate when multiple observed variables (e.g., questionnaire items) are used to measure latent variables (e.g., a person's unobservable level of extroversion). CFA is used to test one or more theories that posit a construct's dimensionality. Some theorists may argue that extroversion is unidimensional. Others may propose that extroversion has two dimensions: extroversion around people one knows and extroversion around strangers. CFA enables both theories to be empirically evaluated using factor models that represent the proposed underlying dimensionality. CFA theoretically corrects for measurement error by using multiple observed variables to measure latent variables. Multiple questionnaire items would be used to measure extraversion and other constructs of interest.

Structural regression models combine path models and CFA by using multiple observed variables to measure latent variables (CFA), and providing estimates of causal relations among these latent variables (path models). Structural regression models usually provide more accurate estimates of causal relations than path models by modeling latent variables using multiple observed variables. Path models assume no measurement error and can underestimate causal effects by not modeling latent variables. Structural regression models are more parsimonious than CFAs. Structural regression models

usually assume some paths/relationships are zero. In CFA, all possible relationships among latent variables or factors are estimated. Thus, CFA is a special case of SEM with no predictive hypotheses among latent variables.

Each of these SEMs has value, but my dissertation primarily addresses CFA and structural regression models. For the rest of the document, SEM will refer to CFA and structural regression models. I will distinguish between CFA and full structural regression models when providing model equations and when defining the specific models estimated in my study. As noted, CFAs and SEMs estimate the relationship among latent variables while correcting for measurement error. These models assume that variability in observed responses can be separated into systematic variability that is shared across items and random variability. Systematic variability is assumed to represent variability due to the theoretical construct (i.e., explained variability). Random variability is assumed to represent measurement error (i.e., unexplained variability). SEMs contain a measurement model defining how the observed variables relate to the latent variable(s) and a structural model defining latent variables' relationships.

The five main estimated parameters in SEM are factor loadings, error variances, factor variances, factor covariances, and structural path coefficients. The underlying constructs or latent variables measured by the observed variables usually are referred to as factors. For each item, SEM has a factor loading and error variance. The factor loading indicates the size of the relationship between the item and underlying factor (i.e., systematic variability). The error variance indicates the amount of variance in the item that is unexplained by the underlying factor (i.e., unexplained variance). Factor variances

indicate the construct/latent variables' variances. If the set of items measure multiple factors, these factors' relationship(s) is estimated with factor covariances. Structural path coefficients estimate the causal relationship among latent variables. CFA provides factor loadings, error variances, factor variances, and factor covariances. Structural regression models contain these parameters and the structural path coefficients.

Multilevel Structural Equation Models

SEM assumes non-nested data. The SEM equations do not account for group membership by allowing parameters to vary based on group membership. Applying SEM to nested data can result in underestimated standard errors, overestimated chi-square values, and biased parameter estimates in either direction (Brown 2015; Julian, 2001; Muthén, 1994; Stapleton, 2013). Thus, Muthén and Asparouhov (2008) introduced the multilevel SEM (MSEM) to explicitly account for nested data. In MSEM, the observed covariance matrix is separated into a Level-1 matrix (i.e., within-groups) and a Level-2 matrix (i.e., between-groups).

The within-groups covariance matrix is a variance/covariance matrix of individual deviations from the group's mean. The between-groups covariance matrix is a variance/covariance matrix of group mean deviations from the mean of all group means. Consider employees nested within managers; employees rate satisfaction with their manager using several items. Individual deviations from the group mean for a given item are employees who indicate low satisfaction with their manager despite high average satisfaction for that manager. Group mean deviations from the mean satisfaction across all group mean satisfaction scores for a given item occur if some managers' mean

satisfaction was low despite high average satisfaction across all managers. Deviations from the mean are based on each individual item, rather than all items at once.

These covariance matrices are modeled separately, yet simultaneously. By modeling these matrices separately, MSEM accounts for the nested data structure. Thus, in situations with nested data, MSEM generally produces more accurate parameter estimates, standard errors, and fit indices than SEM, which models the total covariance matrix (Brown, 2015).

The main difference between SEM and MSEM is that MSEM provides parameter estimates for each level. MSEM provides error variances, factor loadings, factor variances, factor covariances, and regression coefficients for each level. If the models are identical at both levels, the number of estimated parameters is doubled. MSEM enables examination of how parameters differ across levels, which may be of substantive interest (Zyphur, Kaplan, & Christian, 2008). Factor correlations may be much higher at one level, which could result in a different factor structure across levels (e.g., Sessoms & Willse, 2019). If continuous observed variables are analyzed using Maximum Likelihood, separate single-level SEMs could be fit to the within and between matrices. However, this shortcut cannot be used with categorical variables. Categorical estimation methods for MSEM requires analysis of the full data, rather than the summary statistics.

Need for the Current Study

The MSEM literature has important limitations. MSEM simulations generally analyze continuous, normally distributed variables, use small models, and do not study the effect of data asymmetry (skewed/non-normal data). MSEM applications often fit

large models to categorical data that are skewed and asymmetric (Kim et al., 2016). MSEM simulation studies generally also study model fit using Maximum Likelihood estimation. Maximum Likelihood estimation often is not appropriate depending on data characteristics (e.g., number of response categories). Moreover, model fit indices' performance has depended on estimation method (e.g., Yu & Muthén, 2002). The current study will address these limitations. I will conduct a simulation that studies model size, asymmetric/skewed data, partially incorrect/misspecified models, model fit, and alternative estimation methods. In Chapter 2 I will describe in more detail SEM and MSEM, model fit evaluation, common estimation methods, and limitations of MSEM research. I will conclude Chapter 2 with a general description of my research questions.

CHAPTER II

LITERATURE REVIEW

Structural Equation Models (SEM)

SEM refers to path analysis, CFA, and structural regression models. My dissertation focuses on CFA and structural regression models. SEM will refer to CFA and structural regression models for the rest of the document.

SEM (CFA and structural regression models) is used to estimate latent variables' relationships while correcting for measurement error. SEM assumes that variability in observed responses can be separated into systematic variability that is shared across items and random variability. Systematic variability represents variability due to the construct. Random variability represents measurement error. SEMs contain a measurement model and structural model. The measurement model (CFA) defines how the observed indicators (e.g., items) relate to the construct. The structural model (structural regression model) defines predictive relationships among latent variables (e.g., direct effects).

SEM equations. The measurement portion (CFA) of SEM is

$$\mathbf{Y}_i = \boldsymbol{\nu} + \boldsymbol{\Lambda}\boldsymbol{\eta}_i + \mathbf{K}\mathbf{X}_i + \boldsymbol{\varepsilon}_i, \quad (1)$$

where for an individual *i*, *m* latent variables, *p* observed variables, and *q* observed covariates, $\mathbf{Y}_i$ contains *p* observed scores, $\boldsymbol{\nu}$ contains *p* variable intercepts, $\boldsymbol{\Lambda}$ contains *p* factor loadings for *m* latent variables, $\boldsymbol{\eta}_i$ contains *m* latent variable scores, $\mathbf{K}$ contains

$p \times q$ regression coefficients, $\mathbf{X}_i$ contains q observed covariate scores, and $\boldsymbol{\varepsilon}_i$ contains p residual scores (i.e., unexplained variability) with multivariate normality, zero means, and covariance matrix $\boldsymbol{\Theta}$. If only estimating CFA, $\mathbf{K}$ and $\mathbf{X}_i$ drop out given CFA typically excludes regressions except factor loadings.

The structural model of SEM (structural regression model) is

$$\boldsymbol{\eta}_i = \boldsymbol{\alpha} + \mathbf{B}\boldsymbol{\eta}_i + \boldsymbol{\Gamma}\mathbf{X}_i + \boldsymbol{\zeta}_i \ , \tag{2}$$

where for an individual i, $\boldsymbol{\eta}_i$ contains m latent variable scores, $\boldsymbol{\alpha}$ contains m structural intercepts, $\mathbf{B}$ contains regression coefficients for m latent variables regressed onto other m latent variables, $\boldsymbol{\Gamma}$ contains regression coefficients for m latent variables regressed onto q observed covariates, $\mathbf{X}_i$ contains observed scores on q covariates, and $\boldsymbol{\zeta}_i$ contains m residual latent variable scores with multivariate normality, zero means, and covariance matrix $\boldsymbol{\Psi}$.

Model fit in SEM. Methodologists usually advise evaluating model fit before inspecting SEM parameter estimates (Brown, 2015; Kline, 2011). Model fit indicates whether the observed data is predicted accurately based on the model. Good model fit indicates that the observed variances and covariances (and means if modeled) align with their expected values based on the fitted model. Good model fit suggests that parameter estimates are appropriate to interpret because the model's plausibility is supported by the data. Poor model fit suggests that parameter estimates should not be interpreted.

Many model fit indices have been developed. They usually are based on model chi-square, which is sensitive to sample size. The larger the sample, the larger the chi-square; large chi-square values suggest poor fit. With large samples, chi-square can be very large (suggesting substantial misfit) even with minor misfit. Chi-square's influence on fit indices is particularly relevant when fitting MSEMs given typical large Level-1 samples but small Level-2 samples (e.g. 1000 students but 50 teachers). These issues will be discussed further when introducing the importance of level-specific indices.

Multilevel Structural Equation Models (MSEM)

SEM assumes single-level data. Analyzing nested data using SEM can result in biased parameter estimates, standard errors, and fit indices (Brown 2015; Julian, 2001; Muthén, 1994; Stapleton, 2013). MSEM accounts for nested data by providing parameter estimate for each level. In nested data situations, MSEM generally produces more accurate parameter estimates, standard errors, and fit indices than SEM. MSEM also enables inspection of differences in parameter estimates across levels. These differences may be theoretically interesting (Zyphur et al., 2008).

MSEM contains several different types of latent variables. Variability in observed responses (e.g. item responses) is separated into latent within and latent between variability. Common factors are latent variables and refer to the underlying construct measured by the items. MSEM also enables random intercepts and random slopes, which are considered latent variables. Random intercepts refer to intercepts that vary across groups. Random slopes are factor loadings relating a variable to the common factor that vary across groups and structural causal paths relating a common factor to other common

factors that vary across groups. MSEM simulations and applications usually use random intercept MSEMs. Random slopes are uncommon in MSEM simulations and applications. Random slopes show measurement non-invariance; the item's relationship to the common factor (construct) differs based on groups.

MSEM equations. Muthén and Asparouhov (2008) defined MSEM with random intercepts and slopes using:

$$\mathbf{Y}_{ij} = \boldsymbol{\nu}_j + \boldsymbol{\Lambda}_j \boldsymbol{\eta}_{ij} + \mathbf{K}_j \mathbf{X}_{ij} + \boldsymbol{\varepsilon}_{ij} \tag{3}$$

$$\boldsymbol{\eta}_{ij} = \boldsymbol{\alpha}_j + \mathbf{B}_j \boldsymbol{\eta}_{ij} + \boldsymbol{\Gamma}_j \mathbf{X}_{ij} + \boldsymbol{\zeta}_{ij} \tag{4}$$

$$\boldsymbol{\eta}_j = \boldsymbol{\mu} + \boldsymbol{\beta}\boldsymbol{\eta}_j + \boldsymbol{\gamma}\mathbf{X}_j + \boldsymbol{\zeta}_j \ , \tag{5}$$

where *i* is individual and *j* is group. Equation 3 defines the measurement model. $\mathbf{Y}_{ij}$ contains *p* observed variable responses, $\boldsymbol{\nu}_j$ contains *p* observed variable intercepts, and $\boldsymbol{\Lambda}_j$ is $(p \times (2p + 2m))$ and contains factor loadings at each level relating the observed variable to its latent part at each level and to *m* common factors at each level. $\boldsymbol{\Lambda}_j$ is $\left[\mathbf{I}_p \mathbf{0}_{p\times m} \mathbf{I}_p \mathbf{0}_{p\times m}\right]$ where 1 is assigned to each observed variable's path to its latent between variability and its latent within variability, but all other paths are 0. $\boldsymbol{\eta}_{ij}$ contains *p* within-cluster latent parts, *m* within-cluster common factor scores, *p* between-cluster latent parts, and *m* between-cluster common factor scores. $\mathbf{K}_j$ contains regression coefficients for *p* observed variables regressed onto *q* observed covariates, $\mathbf{X}_{ij}$ contains *q* observed covariate scores, and $\boldsymbol{\varepsilon}_{ij}$ contains *p* residuals.

Equation 4 defines the Level-1 structural model. $\boldsymbol{\alpha}_j$ is ($2p + 2m$) and contains p item intercepts and m between-cluster common factor intercepts. $\mathbf{B}_j$ contains within-cluster factor loadings relating the p latent parts to m common factors and predictive paths among Level-1 common factors and is $(2p + 2m)\times(2p + 2m)$. $\boldsymbol{\Gamma}_j$ contains regression coefficients for q observed covariates predicting $(2p + 2m)\times(2p + 2m)$ latent variables. $\boldsymbol{\zeta}_{ij}$ contains within-cluster residuals.

Equation 5 defines the Level-2 structural model. Note, r is the number of random effects and s is the number of cluster-level covariates. $\boldsymbol{\eta}_j$ contains all r random effects from Equations 3 and 4 ($\boldsymbol{\nu}_j$, $\boldsymbol{\alpha}_j$, $\boldsymbol{\Lambda}_j$, $\mathbf{B}_j$, $\mathbf{K}_j$, and $\boldsymbol{\Gamma}_j$). $\boldsymbol{\mu}$, $\boldsymbol{\beta}$, and $\boldsymbol{\gamma}$ contain fixed effects. $\boldsymbol{\mu}$ is ($r\times 1$) and contains the means of random effect distributions and group-level structural intercepts. $\boldsymbol{\beta}$ ($r\times r$) contains Level-2 factor loadings relating the Level-2 variability in item responses to the Level-2 common factors and regression coefficients among Level-2 common factors. $\boldsymbol{\gamma}$ is ($r\times s$) and contains regression coefficients that predict Level-2 factor loadings and/or Level-2 common factors using Level-2 observed covariates. $\boldsymbol{\zeta}_j$ contains r Level-2 residuals with multivariate normality, zero means, and covariance matrix $\boldsymbol{\Psi}$.

My dissertation will fit random intercept MCFAs, which are a special case of MSEM (Geldhof et al, 2014; Sessoms & Willse, 2019). MCFA can be defined by constraining certain elements in Equations 3 through 5. MCFA usually does not use covariates or contain regressions among latent variables except for factor loadings

relating an item's Level-1 or Level-2 variability to its common factor at that level. Excluding covariates drops out observed covariate scores ($\mathbf{X}_{ij}$ and $\mathbf{X}_{j}$) and regression coefficients using these covariates to predict other observed or latent variables ($\mathbf{K}_{j}, \boldsymbol{\Gamma}_{j}$, and $\boldsymbol{\gamma}$) . Measurement model residuals $\boldsymbol{\varepsilon}_{ij}$ drop out because they define prediction error given all factor loadings are fixed to 1 (i.e., no measurement error). Variable intercepts $\boldsymbol{\nu}_{j}$ can be excluded given intercepts are modeled at each level. Level-1 factor loadings are assumed to be constant across groups (i.e., no random slopes), yielding $\boldsymbol{\Lambda}$ and $\mathbf{B}$.

Defining MCFA using these constraints simplifies Equations 3 through 5:

$$\mathbf{Y}_{ij} = \Lambda\eta_{ij} \tag{6}$$

$$\eta_{ij} = \alpha_{j} + \mathbf{B}\eta_{ij} + \zeta_{ij} \tag{7}$$

$$\eta_{j} = \mu + \beta\eta_{ij} + \zeta_{j}, \tag{8}$$

MCFA usually does not contain regression coefficients among common factors. The vector/matrix elements corresponding to these regression coefficients are assumed to be zero (i.e., not estimated). Thus, $\mathbf{B}$ contains Level-1 factor loading estimates, ζ_{ij} contains Level-1 measurement error residual estimates, β contains Level-2 factor loading estimates, and ζ_{j} contains Level-2 measurement error residual estimates. MSEM provides parameter estimates for each level. Single-level SEM ignores this possible source of variation across groups and provides one set of parameter estimates.

SEM can accommodate nested data using multiple-group analyses. However, these multiple-group SEMs provide a limited number of fixed effects and no random effects.

Conceptual example of MSEM and MCFA. A conceptual example of MSEM and MCFA may make these equations more concrete. Consider employees nested within managers. Employees indicate their satisfaction with their manager by completing a questionnaire. MCFA would be appropriate if researchers only were interested in testing hypothesized dimensionality. Hypothesized dimensionality would be based on construct theories and previous research. Theory A might posit that satisfaction with manager is unidimensional at both levels. Theory B could argue that satisfaction with manager is two-dimensional at both levels. These two distinct aspects of satisfaction could be 1) satisfaction with how the manager treats the respondent and 2) satisfaction with how the manager treats other employees. MCFA would test these two hypotheses about construct dimensionality to evaluate which has more support from the data. Both random intercepts and random slopes are possible. Random slope MCFAs usually are not tested given implied measurement non-invariance of Level-1 factor loadings (i.e., item-factor relationship differs across managers). Random intercepts MCFA is much more common.

MSEM would be appropriate if the researchers also wanted to predict employees' satisfaction with managers. Predictors could be added at the employee level, manager level, or both. Employee-level predictors could be employees' satisfaction with interpersonal relationships (e.g. family, friends). Manager-level predictors could be manager expectations about work-life balance. Each predictor would be modeled using multiple indicators (i.e., multiple questionnaire items). $\mathbf{B}$ would contain the predictive

relationship between employee satisfaction with interpersonal relations and employee satisfaction with manager. $\mathbf{B}$ also would contain the Level-1 factor loadings relating Level-1 variability in items measuring employee satisfaction with managers to the Level-1 construct/common factor for employee satisfaction with managers. ζ_{ij} would contain estimated prediction errors based on using employee satisfaction with interpersonal relationships to predict employee satisfaction with manager. ζ_{ij} also would contain Level-1 measurement error residuals, which indicate some Level-1 variability in the items is not explained by the Level-1 common factor. β would contain the predictive relationship between manager's work-life expectations and employee satisfaction with manager. β also would contain Level-2 factor loadings. ζ_j would contain estimated prediction errors based on using manager work-life expectations to predict employee satisfaction with manager. ζ_j also would contain Level-2 measurement error residuals. MSEM typically only uses latent predictors, dropping out observed covariate scores ($\mathbf{X}_{ij}$ and $\mathbf{X}_j$) and their regression coefficients ($\mathbf{K}_j, \mathbf{\Gamma}_j, \gamma$). This approach yields Equations 6 through 8 that previously defined the MCFA. However, in MCFA, covariate regression coefficients in $\mathbf{B}$ and β are assumed to be zero (i.e., not estimated). With MCFA, $\mathbf{B}$ and β only contain estimated values for level-specific factor loadings. In MCFA, prediction errors in ζ_{ij} and ζ_j are not estimated but fixed to zero given no predictors. Instead, in MCFA, ζ_{ij} and ζ_j only contain estimates of measurement error residuals.

MSEM Literature Overview

The MSEM literature is vast. Most MSEM research evaluates how sample size, the degree of nesting in the data, and/or estimation method affects parameter estimation. Some MSEM research extends single-level reliability and fit indices to MSEM (Geldhof et al., 2014; Ryu & West, 2009). Level-specific reliability and fit indices' accuracy has been explored (Boulton, 2011; Geldhof et al., 2014; Hsu et al., 2017; Ryu & West, 2009; Yuan & Bentler, 2007).

Intraclass correlation coefficients (ICCs), sample size, model fit, and estimation in MSEM will be discussed next. The degree of nesting in the data (i.e., ICCs) will be discussed because it is frequently studied in MSEM and impacts these three issues. Appropriate ICCs, sample sizes, fit indices, and estimation method are interdependent. Thus, advice for MSEM applications often discusses these four issues in tandem.

Intraclass correlation coefficient (ICC). The ICC, also known as rho, is the proportion of variance in the variable (e.g., student perceptions of his/her teacher) that is attributable to between-group differences (e.g., differences in average teacher perceptions between two teachers). ICCs, sometimes referred to as rho in the multilevel literature (e.g., Raudenbush & Bryk, 2002) are estimated using:

$$\rho = \frac{\sigma_B^2}{\sigma_B^2 + \sigma_W^2} \quad , \tag{9}$$

where σ_B^2 is the proportion of variance due to between-group differences and σ_W^2 is the proportion of variance due to within group differences (e.g., differences between two

students' perception of the same teacher). ICCs are important to examine because if ICCs are close to zero, MSEM may be unnecessary (Brown, 2015). Low ICCs indicate that most variability is occurring within groups, not between groups. For example, most variability in perceptions of teachers is due to differences among students, not differences among teachers. Low ICCs also can cause model non-convergence (Clifton & Depaoli, 2017; Depaoli & Clifton, 2015; Kim et al., 2012), bias Level-2 parameter estimates (Ludtke et al., 2008), and reduce fit indices' power to detect Level-2 misfit (Boulton, 2011; Hsu et al., 2017). ICCs may be more important for convergence than group size and number of groups (DePaoli & Clifton, 2015; Hsu et al., 2017).

A latent factor ICC also has been proposed for MSEM (e.g., Hsu et al., 2017; Muthén, 1994). The latent factor ICC is analogous to the observed variable ICC except that it represents the proportion of factor variance due to between-groups:

$$ICC_{\eta} = \frac{\psi_B}{\psi_B + \psi_W} \quad , \tag{10}$$

where ψ_B represents the proportion of factor variance at the between level and ψ_W represents the proportion of factor variance at the within level. The latent factor ICC requires the same factor structure across levels.

Methodologists often disagree on the minimum acceptable ICC values. The acceptability of ICCs also usually depends on the number of groups and group size. Generally, appropriate observed ICC values are discussed more than latent factor ICCs. Acceptable ICC values have fluctuated depending on the specific fit index used (Hsu et

al., 2017). Hsu et al. (2017) generally advised latent ICC values of at least .09, .17, or .23 depending on the fit index used for assessing between-level fit. Kim et al. (2012) suggested latent ICC values exceeding .33 when performing multilevel group invariance testing. Observed ICCs > .05 have been suggested for MSEM (Brown, 2015; Geldhof et al., 2014; Julian, 2001). Preacher et al. (2011) advised ICCs ≥ .10. ICCs < .10 or .05 can cause MSEM non-convergence (e.g., Boulton, 2011; Clifton & Depaoli 2017; Depaoli & Clifton, 2015). Hox (2010) suggested that ICCs ≥ .15 generally would be considered large in most applied settings. Observed ICCs > .30 generally are rare in application (Ludtke et al., 2008). However, a review of multilevel factor analysis applications found observed ICCs ranged from .04 to .70 (Kim et al. 2016). Some MSEM simulation studies choose ICC conditions that range from .05 to .15 (e.g., Hox et al., 2010) or .05 to .30 (e.g., Boulton, 2011; Ludtke et al., 2008; Ludtke et al., 2011.) Often MSEM simulation studies use ICCs as high as .40 (e.g., Depaoli & Clifton, 2015; Preacher et al., 2011).

Generally, as model complexity increases, the appropriate ICC also increases. Unfortunately, almost all MSEM research that has studied ICCs has used very simple models (1-2 factors at each level). Thus, higher ICCs may be necessary than common recommended values when complex MSEMs are used. For example, multilevel factor analysis applications often model 4 to 7 factors at each level (Kim et al., 2016).

In conclusion, I generally recommend observed ICCs ≥ .10 when estimating relatively simple MSEMs (e.g., 1-2 factors at each level). ICCs of at least .20 to .25 may be necessary when estimating more complex MSEMs.

Sample size in MSEM. There are three types of sample sizes in MSEM: Level-1 sample size (e.g., total number of students), Level-2 sample size or number of groups (e.g., number of teachers), and group size (e.g., number of students who share the same teacher). Generally, Level-1 sample sizes are sufficient for MSEM and are rarely discussed. Level-1 $N > 2{,}000$ is very common in applications (Kim et al., 2016). Thus, I will only discuss Level-2 sample size (i.e., number of groups) and group size.

Number of groups guidelines. Level-2 sample size represents the number of groups (e.g., number of teachers) and often is considered more important than Level-1 sample size (e.g., Raudenbush & Bryk, 2002). Unfortunately, Level-2 sample sizes usually are much smaller than Level-1 (e.g., 100 teachers compared to 2,000 students). Small Level-2 sample sizes can cause non-convergence and overestimation of factor loadings (e.g., Depoli & Clifton, 2015). Number of groups sometimes affects parameter estimation accuracy more than ICCs, group size, or number of estimable parameters (e.g., Hox & Maas, 2001).

Appropriate Level-2 sample size can depend on the estimation method. Maximum Likelihood (ML) estimation typically requires larger samples than Bayesian estimation. Bayesian estimation has performed accurately with as few as 20 groups (Hox, van de Schoot, & Matthijsse, 2012). ML has performed poorly with 20 groups (Hox et al., 2012; Meuleman & Billiet, 2009). Hox et al. (2012) advised 40-60 groups when using ML. Hox et al. (2010) suggested 50 groups when using ML or Diagonally Weighted Least Squares estimation.

The appropriate number of groups also may depend on model complexity. Meuleman and Billiet (2009) suggested 40 groups for multilevel CFA (MCFA). MSEM has overestimated Level-2 covariate effects (Depaoli & Clifton, 2015) and factor loadings (Hox & Maas, 2001; Preacher et al. 2011) with < 100 groups. However, this bias disappeared when corresponding loadings were constrained to equality (e.g., Item 1's loading fixed to equality at Level-1 and Level-2; Depaoli & Clifton, 2015). They found that 40 groups provided accurate estimates with corresponding loadings fixed to equality, ICCs ≥ .10, and group size > 10. These equality constraints, which reduce model complexity, may increase convergence and accuracy even with low ICCs and few groups if parameter estimates are similar (Depaoli & Clifton, 2015; Lüdtke, Marsh, Robitzsch, & Trautwein, 2011; Meuleman & Billiet, 2009).

Unfortunately, most of these recommendations are based on MSEM simulation studies that use very simple models: 1-2 factors at each level. Conversely, applied MSEM researchers often estimate 4 to 7 factors at each level (Kim et al., 2016). From personal experience, more between-level parameters than groups often result in non-convergence. Even if the model converges, it also often yields severely underestimated factor loadings and standard errors and overestimated structural effects (Meuleman & Billiet, 2009). Thus, applied researchers estimating more than 2 factors at each level may need more groups.

In conclusion, I generally recommend 40-50 groups when estimating relatively simple MSEMs (e.g., 1-2 factors per level). Conversely, at least 60-100 groups likely are needed when estimating complex MSEMs (e.g., 4 factors per level).

Group size guidelines. Group size refers to the number of Level-1 units within each Level-2 unit (e.g., the number of students who share a teacher). Methodologists sometimes disagree whether group size is as important as ICCs and number of groups. Many groups with few members may be preferable to few groups with many members (Meuleman and Billiet, 2009). For parameter estimation accuracy, large group sizes have compensated partially for small ICCs and few groups (Hox & Maas, 2001). Sessoms and Willse (2019) noted that group sizes of 5 was common advise when Level-1 parameter estimates were primary interest. Preacher et al. (2011) advised group sizes of 20 for MSEM. Cheung and Au (2005) suggested group sizes of 50 or 100 for MSEM. Depaoli Clifton (2015) advised group sizes ≥ 10 when ICCs $\geq .10$ and groups ≥ 40. As with ICCs and number of groups, increasing group size usually increases results' accuracy.

I generally recommend group sizes of 20 to 50. Group sizes closer to 50 likely are necessary when estimating relatively complex MSEMs (e.g., 4 factors at each level). Group sizes closer to 20 may be appropriate when estimating relatively simple MSEMs (e.g., 1-2 factors at each level).

Model Fit Evaluation in MSEM: Importance of Level-Specific Fit

MSEM fit evaluation is complex. Applied researchers usually evaluate MSEM fit by estimating the full model (expected factor structure and causal paths at both levels) and inspecting the aggregate fit indices that combine fit for both levels (Kim et al., 2016). In contrast, methodologists generally recommend level-specific fit (Ryu & West, 2009; Stapleton, 2013; Yuan & Bentler, 2007). Level-specific fit addresses fit separately for each level. Applied researchers may tend to report aggregate fit indices partially because

level-specific fit indices are not automatically computed in most MSEM software programs (Sessoms & Willse, 2019).

Aggregate fit evaluation is problematic for two reasons. First, if the full MSEM fits poorly, misfit may be due to the within level, between level, or both (Ryu & West, 2009). Second, aggregate fit indices likely reflect Level-1 (mis)fit given a much larger Level-1 sample. Recall that most fit indices are based on model chi-square, which is very sensitive to sample size. With large samples, chi-square will be very large (suggesting substantial misfit) even with minor misfit. Most MSEM applications have much larger Level-1 samples than Level-2 samples (e.g. 1000 students and 50 teachers). Aggregate fit indices combine Level-1 (mis)fit and Level-2 (mis)fit. Thus, aggregate fit indices will reflect primarily Level-1 (mis)fit and mask Level-2 (mis)fit.

Level-specific fit usually outperforms simultaneous evaluation of fit (Boulton, 2011; Ryu & West, 2009; Sessoms & Willse, 2019; Yuan & Bentler 2007). This multiple-step approach to evaluating fit is analogous to the two-step process in structural equation modeling (SEM). This process fits the measurement model first, then tests the full structural model (Kline, 2011). The same logic applies to both approaches: separately estimating each step better identifies misfit.

Yuan and Bentler's Approach to Level-Specific Fit

Two general approaches to level-specific fit have been suggested. Yuan and Bentler (2007) introduced the segregating approach that estimates the within and between-group covariance matrices. Yuan and Bentler's (2007) approach then fits a model to each matrix one at a time using conventional (non-multilevel) SEM software.

One could fit 3 factors to the within covariance matrix only and obtain fit results. Next, one would fit 3 factors to the between covariance matrix and obtain fit results. However, Yuan and Bentler's (2007) approach reduces the possible estimation methods that can be applied. Estimation methods that require analyzing the raw data instead of analyzing the covariance matrix (e.g., Weighted Least Squares estimation) cannot be used. These estimation methods that require raw data for analysis often are most appropriate in many contexts (e.g., categorical data). The proper estimation method depends on the data; incorrect choices can severely bias results (Finney et al., 2016).

Ryu and West's Approach to Level-Specific Fit

Ryu and West (2009) introduced the partially-saturated approach to assessing level-specific fit. This approach estimates baseline/independence models (all level-specific variables uncorrelated) separately for each level. Next, one estimates a theoretical model (expected number of factors) at each level separately. If the model fits well at each level separately, then the full theoretical model is fit to both levels (Stapleton, 2013). When estimating the baseline and theoretical models at each level, a saturated model is fit to the other level that freely estimates all covariances at that level. By freely estimating all covariances at a given level, that model will fit perfectly and thus cannot affect misfit. Any occurring misfit can only be due to the other level. Estimating the baseline and theoretical models are necessary to compute MSEM fit indices. For example, the Comparative Fit Index is calculated using the chi-square and degrees of freedom from the independence model and theoretical model. Equations for level-specific fit indices are identical to their traditional single counterparts. Both level-specific

approaches have outperformed aggregate fit (Boulton 2011; Ryu & West, 2009; Yuan & Bentler, 2007). The two level-specific fit methods generally provide similar fit index values except when ICCs are low (Boulton, 2011). Both fit approaches performed poorly with ICCs ≤ .05 (Boulton, 2011).

Defining level-specific fit indices. When performing MSEM, fit should be evaluated separately for each level to accurately identify the source of misfit. If the full MSEM fits poorly, misfit may be due to the within level, between level, or both. Ryu and West (2009) and Muthén and Muthén (1998-2017) defined level-specific fit equations. Ryu and West (2009) adapted the Comparative Fit Index (CFI) and Root Mean Square Error of Approximation (RMSEA). Muthén and Muthén (1998-2017) adapted the Standardized Root Mean Square Residual. These indices were adapted because they performed best in previous evaluations of fit indices (e.g., Hu & Bentler, 1998, 1999). The equations for level-specific CFI, TLI, RMSEA, and SRMR are identical to their traditional single-level counterparts.

The chi-square statistic (χ^2) used to compute most fit indices is defined using:

$$\chi^2 = F(N-1), \tag{11}$$

where N is the sample size, and F is the minimum fit function applied by the estimation method that seeks to minimize the difference between the observed covariance or correlation matrix and the model-predicted covariance matrix. χ^2 accounts for clustered data because it is estimated using MSEM, which models data as clustered.

Degrees of freedom (*df*) in SEM is defined using:

$$df = \frac{p(p+1)}{2} - r \ , \tag{12}$$

where p is the number of variables and r is the number of estimated parameters. Sample size is not involved in computing degrees of freedom. Thus, degrees of freedom are a function of the number of variables and estimated parameters but not sample size.

The chi-square statistic is used to compute these level-specific fit indices. The level-specific chi-square accounts for the clustering. If the full theoretical model were fit at both levels (e.g., 2 factors at both levels), the aggregate chi-square would reflect misfit due to both levels. For example, assume that the full model contains misfit at both levels. Thus, the aggregate χ^2 could be 200, and

Level-specific CFI. The adapted CFI for Level-1 is:

$$\text{Ll CFI} = 1 - \frac{Max[\chi^2_{PSw} - df_{PSw}, 0]}{Max[\chi^2_{PSiw} - df_{PSiw}, 0]} \tag{13}$$

where χ^2_{PSw} and df_{PSw} are the χ^2 and *df* associated with the Level-1 theoretical model that estimates the Level-1 theoretical model and freely estimates all Level-2 covariances (i.e., fully saturated Level-2 model), and χ^2_{PSiw} and df_{PSiw} are the χ^2 and *df* associated with the Level-1 independence model that fixes all Level-1 covariances to zero and freely estimates all Level-2 covariances. Freely estimating all Level-2 covariances for both

models isolates Level-1 fit. The level with freely estimated covariances cannot affect misfit because covariances are reproduced perfectly.

The adapted CFI for Level-2 is:

$$\text{L2 CFI} = 1 - \frac{Max[\chi^2_{PSb} - df_{PSb}, 0]}{Max[\chi^2_{PSib} - df_{PSib}, 0]} \tag{14}$$

where χ^2_{PSb} and df_{PSb} are the χ^2 and *df* associated with the Level-2 theoretical model that freely estimates all Level-1 covariances and estimates the Level-2 theoretical model, and χ^2_{PSib} and df_{PSib} are the χ^2 and *df* associated with the Level-2 independence model that freely estimates all Level-1 covariances and fixes all Level-2 covariances to zero. Freely estimating Level-1 covariances for both models isolates Level-2 fit. The level with freely estimated covariances cannot affect misfit because covariances are reproduced perfectly.

Level-specific TLI. Level-1 TLI is defined using

$$\frac{\left(\chi^2_{PSiw} / df_{PSiw}\right) - \left(\chi^2_{PSw} / df_{PSw}\right)}{\left(\chi^2_{PSiw} / df_{PSiw}\right)} \tag{15}$$

where χ^2_{PSw} and df_{PSw} are the χ^2 and *df* associated with the Level-1 theoretical model that estimates the Level-1 theoretical model and freely estimates all Level-2 covariances (i.e., fully saturated Level-2 model), and χ^2_{PSiw} and df_{PSiw} are the χ^2 and *df* associated with the Level-1 independence model that fixes all Level-1 covariances to zero and estimates all Level-2 covariances. These elements are identical to those in the Level-1 CFI equation.

Level-2 TLI is defined using

$$\frac{\left(\chi^2_{PSib} / df_{PSib}\right) - \left(\chi^2_{PSb} / df_{PSb}\right)}{\left(\chi^2_{PSib} / df_{PSib}\right)} \tag{16}$$

where χ^2_{PSb} and df_{PSb} are the χ^2 and *df* associated with the Level-2 theoretical model that freely estimates all Level-1 covariances and estimates the Level-2 theoretical model, and χ^2_{PSib} and df_{PSib} are the χ^2 and *df* associated with the Level-2 independence model that freely estimates all Level-1 covariances and fixes all Level-2 covariances to zero. These elements are identical to those in the Level-2 CFI equation.

Level-specific RMSEA. The adapted RMSEA for Level-1 fit is:

$$\text{L1 RMSEA} = \sqrt{\frac{\chi^2_{PSw} - df_{PSw}}{df_{PSw}(N)}} \tag{17}$$

where χ^2_{PSw} and df_{PSw} are the χ^2 and *df* for the simultaneously estimated theoretical model at Level 1 and fully saturated model at Level-2, and *N* is the Level-1 sample size.

The adapted RMSEA for Level-2 is:

$$\text{L2 RMSEA} = \sqrt{\frac{\chi^2_{PSb} - df_{PSb}}{df_{PSb}(J)}} \tag{18}$$

where χ^2_{PSb} and df_{PSb} are the χ^2 and *df* for the simultaneously estimated fully saturated model at Level 1 and theoretical model at Level 2, and *J* is the Level-2 sample size.

Level-specific SRMR. The adapted SRMR for Level-1 is:

$$\text{L1 SRMR} = \sqrt{\frac{\left\{2\sum_{x_w=1}^{p}\sum_{y_w=1}^{x_w}[(s_{x_w y_w} - \hat{\sigma}_{x_w y_w})/(s_{x_w x_w} s_{y_w y_w})]^2\right\}}{p(p+1)}}, \tag{19}$$

where $s_{x_w y_w}$ is the sample covariance of within-level variables x_w and y_w, $\hat{\sigma}_{x_w y_w}$ is the model-implied covariance of within-level variables x_w and y_w, p is the number of observed variables, $s_{x_w x_w}$ is the standard deviation of x_w, and $s_{y_w y_w}$ is the standard deviation of y_w.

The adapted SRMR for Level-2 is:

$$\text{L2 SRMR} = \sqrt{\frac{\left\{2\sum_{x_b=1}^{p}\sum_{y_b=1}^{x_b}[(s_{x_b y_b} - \hat{\sigma}_{x_b y_b})/(s_{x_b x_b} s_{y_b y_b})]^2\right\}}{p(p+1)}}, \tag{20}$$

where $s_{x_b y_b}$ is the sample covariance of between-level variables x_b and y_b, $\hat{\sigma}_{x_b y_b}$ is the model-implied covariance of between-level variables x_b and y_b, p is the number of observed variables, $s_{x_b x_b}$ is the standard deviation of x_b, and $s_{y_b y_b}$ is the standard deviation of y_b.

Aggregate vs. Level-Specific Fit in Application

Aggregate fit indices primarily reflect Level-1 fit given a much larger sample size at Level-1 than Level-2. Sessoms and Willse (2019) didactically demonstrated this dominance of Level-1 fit when assessing aggregate fit indices. Their data were 2,827 students (Level-1 sample size) and 62 teachers (Level-2 sample size). The much larger sample size at Level-1 than Level-2 suggested that the aggregate fit indices would primarily reflect Level-1 fit. They fit a full MSEM that specified completely uncorrelated variables at Level-2 and freely estimated all Level-1 covariances (perfect Level-1 fit). This Level-2 model was wrong given the high observed correlations among the variables at Level-2 (*r*'s = .8-.9). The aggregate fit indices still suggested excellent fit (CFI = .99, RMSEA = .04) despite the Level-2 model being completely wrong. In contrast, the level-specific fit indices clearly indicated Level-2 misfit (L2 CFI = .19, L2 RMSEA = .28, L2 SRMR = .69).

Level-Specific Fit Indices' Performance

Level-specific CFI and RMSEA's performance is mixed. Their accuracy usually depends on ICC levels; as ICC increases, their accuracy increases (Boulton, 2011; Hsu et al., 2017). Ryu and West (2009) found that level-specific CFI and RMSEA accurately identified misspecification, but they only used large ICC values of .50. When using the conventional cutoff values proposed by Hu and Bentler (1999), level-specific CFI still accurately identified Level-2 misfit even with low latent ICC values (.09; Hsu et al., 2017). Level-specific RMSEA or SRMR were not recommended under low latent ICC conditions (Hsu et al., 2017). Boulton (2011) found that level-specific CFI, RMSEA, and

SRMR generally were sensitive to model misspecification. This sensitivity decreased as ICC decreased. Note that these evaluations of level-specific fit indices used Maximum Likelihood estimation. When analyzing single-level SEMs, fit indices' performance has differed when using different estimation methods (Yu & Muthén, 2002). Level-specific fit index performance when using different estimation methods is an important, but limited area of research.

MSEM Estimation Methods

The method chosen to estimate MSEMs is important. The proper estimation method depends on the data (e.g., number of response categories, degree of asymmetry). Incorrect choice of estimation method can yield extremely biased results (e.g., incorrect factor loadings estimates, fit indices; Bandalos, 2014; Finney & DiStefano, 2013).

The three most popular estimation methods in MSEM applications are Bayesian, Maximum Likelihood (ML), and Weighted Least Squares (WLS) estimation. Only a basic overview of Bayesian and ML estimation will be provided. I will describe WLS estimation in detail given its value in analyzing attitudinal measures and relative unknown performance in MSEM. Moreover, performance comparisons of WLS to other estimation methods will be presented.

Bayesian estimation. Bayesian estimation of MSEMs has been recommended instead of ML or WLS estimation (Depaoli & Clifton, 2015). Bayesian estimation of MSEMs can yield greater convergence rates than ML and WLS even with few groups and low ICCs (Clifton & Depaoli, 2017; Depaoli & Clifton, 2015). Bayesian estimation uses additional prior information ("priors") to influence parameter estimates. Parameters

are considered to be drawn from a distribution; the user specifies the characteristics of a distribution. The parameter estimates then are influenced based on how likely they are given the distribution from which they were theoretically sampled. Choice of priors will influence the degree of impact on the parameter estimates. Weakly informative priors usually are recommended (Depaoli & Clifton, 2015; Hox et al., 2012; Muthén & Muthén, 1998-2017). Usually, MSEM regression parameters (e.g., factor loadings) are drawn from a multivariate normal prior distribution (Clifton & Depaoli, 2016). MSEM variance parameters typically are drawn from an inverse-gamma prior distribution. The uniform distribution often is used when desiring noninformative priors for cluster-level variances. Truncated-*t* or Cauchy distributions often are used when desiring weakly informative priors for cluster-level variances (Clifton & Depaoli, 2016).

Bayesian estimation is difficult to properly implement (Depaoli & van de Schoot 2016). Bayesian estimation introduces considerable complexity in that different prior values are appropriate depending on the response scale. Priors can meaningfully impact results even when those priors are intended to be weakly informative. Moreover, the appropriate prior distribution depends on the level and specific parameter estimated (Clifton & Depaoli, 2017). Not surprisingly, Bayesian estimation of MSEMs does not always outperform ML or DWLS (e.g., Holtmann et al. 2016). Bayesian estimation is a promising MSEM estimation method, but further study of prior values and distributions is needed. Using Bayesian estimation of MSEMs introduces many complexities outside this manuscript's scope. Thus, further discussion of Bayesian estimation will be limited.

Maximum Likelihood (ML). ML seeks "to identify the population parameters that have the highest probability of producing the sample data" (Enders, 2010, p.84). ML addresses this goal by maximizing the likelihood function. The likelihood function defines the likelihood of a set of parameter estimates and indicates fit between the data and the set of parameter estimates. ML iteratively compares the likelihood function values for several possible sets of parameter values and chooses the set of parameters that is most plausible relative to the other sets of parameter values (Enders, 2010).

ML is the most popular estimation method in MSEM applications (Depaoli & Clifton, 2015; Kim et al., 2016). Methodologists (e.g., Depaoli & Clifton, 2015) usually advise ML when multilevel data are continuous and ML assumptions are met. ML assumes normally distributed data (Kline, 2011). ML generally outperforms other frequentist estimation methods for MSEM as sample size increases and ML assumptions are met (e.g., Hox et al., 2010). However, ML may yield non-convergence and biased parameter estimates when analyzing categorical and/or high dimensional data (Depaoli & Clifton, 2015). Categorical and high-dimensional data are common in MSEM applications (Kim et al., 2016). Multilevel data with high dimensionality generally is extremely difficult to estimate using ML. These nonconvergence problems occur because ML uses one dimension of integration for each factor or random effect (Asparouhov & Muthén, 2007). MSEMs that contain random effects and/or require five or more dimensions to integrate generally do not converge when using ML (Asparouhov & Muthén, 2007; Depaoli & Clifton, 2015). MSEM requires other estimation methods that allow for categorical, high-dimensional data.

Diagonally Weighted Least Squares (DWLS). DWLS is a relatively popular alternative for analyzing multilevel categorical data with high dimensionality. DWLS makes no distributional assumptions. DWLS also does not require high dimensional numerical integration to estimate MSEMs. DWLS instead estimates multiple simple models that each use one or two-dimensional integration (Asparouhov & Muthén, 2007). DWLS' lack of high dimensional numerical integration enables modeling of high dimensional data (Depaoli & Clifton, 2015). Moreover, DWLS computation time is not affected by the number of random effects (Asparouhov & Muthén, 2007).

Methodologists generally advise DWLS for single-level SEM when analyzing fewer than five categories (Bandalos, 2014; Li, 2016). When analyzing fewer than five response categories, ML often underestimates parameters (e.g., factor loadings; Finney, DiStefano, & Kopp, 2016) and sometimes overestimates standard errors (Bandalos, 2014). Performance evaluations of DWLS for MSEM estimation are limited but promising (Asparohouv & Muthén 2007; Hox et al. 2010).

Mathematical Definition of WLS and DWLS Estimation

WLS applies the fitting function

$$F_{WLS} = [\mathbf{s} - \boldsymbol{\sigma}(\boldsymbol{\theta})]' \mathbf{W}^{-1} [\mathbf{s} - \boldsymbol{\sigma}(\boldsymbol{\theta})], \tag{21}$$

where $\mathbf{s}$ is a vector of sample polychoric correlations, $\boldsymbol{\sigma}(\boldsymbol{\theta})$ is the model-implied vector of population polychoric correlations in the matrix $\sum(\theta)$, $\mathbf{W}$ is a positive-definite weight matrix, and $\mathbf{s} - \boldsymbol{\sigma}(\boldsymbol{\theta})$ is the residual vector of the discrepancies between the sample and

model-implied correlations. WLS inverts the weight matrix **W** during estimation. WLS uses the much larger asymptotic covariance matrix as the weight matrix, not the sample covariance matrix as in ML. Thus, WLS estimation is much more computationally intensive. WLS estimation often does not converge with small samples or many indicators (e.g., Flora & Curran, 2004). Because MSEM doubles the number of indicators by analyzing two covariance matrices (between and within) rather than one (total), these problems likely are more apparent in MSEM.

DWLS, also referred to as robust WLS or WLSMV, was created as an alternative to WLS. When estimating parameters, DWLS replaces the full weight matrix **W** with a diagonal matrix **V**. **V** contains the asymptotic variances of the thresholds and polychoric correlation estimates. **V** excludes the covariances of the thresholds and polychoric correlation estimates. Using only some of the available information from the matrix (i.e., variances but not covariances) biases chi-square and standard errors (Savalei, 2014). DWLS applies robust corrections to the chi-square and standard errors to address this bias. These corrections are similar to ML Satorra-Bentler adjustments for non-normality. DWLS corrections involve **W** when estimating the chi-square and parameter standard errors, but do not invert **W.** Inverting **V** instead of **W** (Muthén, du Toit, & Spisic, 1997) improves WLS convergence when analyzing small samples and many variables (Flora & Curran, 2004). **W** also is often nonpositive definite (Flora & Curran, 2004), which prevents inversion when applying the fitting function F_{WLS} .

DWLS Performance

Most evaluations of DWLS performance have estimated single-level SEMs. DWLS performance when analyzing single-level SEMs will be discussed first to inform expectations about DWLS performance when estimating MSEMs. DWLS performance comparisons to WLS and ML will be detailed. Finally, DWLS performance evaluations in the context of MSEM will be described.

Relative performance of DWLS to WLS with single-level SEMs. DWLS estimation generally outperforms WLS estimation when categorical data is analyzed. DWLS typically yields accurate χ^2 values even when analyzing small samples and many variables (Bandalos, 2014; Flora & Curran, 2004; Yang-Wallentin, Jöreskog, & Luo, 2010). DWLS has yielded essentially unbiased and more accurate parameter estimates and standard errors than WLS (Flora & Curran, 2004; Oranje, 2003; Yang-Wallentin et al., 2010). The general accuracy of DWLS fit indices are mixed (Bandalos, 2008; Yu & Muthén 2002).

Relative performance of DWLS to ML with single-level SEMs. DWLS also generally outperforms ML when analyzing categorical data. DWLS tends to yield more accurate factor loadings and factor covariances for most data symmetry conditions (Bandalos, 2014; Finney et al., 2016; Li, 2016) and for both misspecified and correctly specified models (e.g., Bandalos, 2014; Li, 2016). Despite the general preference for DWLS, robust ML sometimes estimates parameters as well under certain conditions (Lei, 2009; Lei & Wu, 2012). Comparisons of standard errors and chi-square estimates are mixed (Bandalos, 2014; Lei, 2009; Li, 2016). However, DWLS usually yields more

accurate fit indices (Finney et al., 2016) and greater power to detect omitted paths (Lei, 2009), weak paths (Li, 2016), omitted factor loadings, and collapsed factors (Bandalos, 2014).

DWLS performance with MSEM. Evaluations of DWLS performance when estimating MSEMs generally are limited. Most MSEM simulation studies use ML estimation. DWLS performance investigations when estimating MSEMs generally study the effect of varying ICC, number of groups, and group size. This research usually evaluates the effects on convergence, Type I and II error, and bias of parameter estimates and standard errors. Little MSEM research has evaluated the accuracy of level-specific fit indices when using DWLS. Increasing ICC, number of groups, and/or group size typically increases convergence and accuracy of results (e.g., Depaoli & Clifton, 2015).

The effect of ICCs on DWLS performance generally varies based on the number of response categories. When analyzing seven categories, ICCs of .05 and .15 did not affect DWLS parameter estimates, standard errors, or model fit indices (Hox et al., 2010). When analyzing dichotomous variables, DWLS resulted in poor convergence and biased parameter estimates except with ICCs of .40 (Depaoli & Clifton, 2015). ICCs $\geq .10$ yielded high convergence and accurate parameter estimates with number of groups ≥ 40, group size ≥ 10, and corresponding loadings constrained to equality (e.g., Item 1's loading fixed to equality at Level-1 and Level-2; Depaoli & Clifton, 2015). These equality constraints reduced model complexity. When analyzing categorical variables with 4 response categories, DWLS convergence, parameter estimates and standard errors generally were accurate when using fixed ICCs of .20 (Holtmann et al., 2016). These

results suggest that ICCs and number of response categories influence DWLS performance in MSEM. Increasing the number of response categories seems to reduce the ICC values needed for convergence and accuracy. Model complexity also seems to affect convergence and accuracy. More complex models may result in non-convergence and biased parameter estimates.

The appropriate number of groups and group members generally depend on each other and the number of response categories. Analyzing dichotomous items and freely estimating all factor loadings caused severe bias and low convergence with 40 or 50 groups and group sizes of 5, 10, and 20 (Depaoli & Clifton, 2015). Constraining corresponding loadings to equality generally produced acceptable convergence and parameter estimation accuracy with ≥ 40 groups and ≥ 10 group members. Analyzing five response categories typically produced unbiased parameter estimates with 100 groups and group sizes of 10; no other sample size conditions were evaluated (Asparohouv & Muthén, 2007). Analyzing seven response categories yielded accurate parameter estimates with 50, 100, and 150 groups (Hox et al., 2010). Increasing group size from 5 to 15 to 25 slightly improved accuracy (Hox et al., 2010). Very small group sizes of 2 often prevented computation of standard errors regardless of the number of groups (50, 100, 150, or 200; Holtmann et al., 2016). Analyzing four response categories using ≥100 groups and group sizes of ≥ 4 generally produced high convergence rates and accurate parameter estimates (Holtmann et al., 2016).

MSEM Research Limitations on DWLS Performance

Model misspecification, model size, and data asymmetry/skewness have received limited attention when studying DWLS in MSEM. These areas have received more attention in single-level SEM studies than MSEM studies. Thus, I will note DWLS performance in single-level studies first for each area to inform expectations about MSEM. I will describe the limitations of MSEM studies related to these three areas and how future research can address these limitations. Discussion of single-level performance will inform expectations about DWLS performance in MSEM scenarios.

Model size. Model size indicates the total number of observed variables and the number of modeled factors. Model size increases as the number of variables increase or the number of modeled factors increase. Model size is important to evaluate because as model size increases, both non-convergence and parameter estimation bias may increase.

Single-level simulation studies evaluating DWLS using medium to large models are limited but promising (Bandalos, 2014; Beauducel & Herzberg, 2006; Li, 2016). DWLS yielded accurate factor loadings regardless of model size; 1, 2, 4, and 8 factors were modeled (Beauducel & Herzberg, 2006). Although model size was not varied, DWLS yielded essentially unbiased factor loadings, factor correlations, and more accurate chi-square than ML when fitting a 5-factor structural model (Li, 2016). DWLS also provided more accurate factor loadings, factor covariances, and most structural paths than ML when analyzing 3 and 7 factors (Bandalos, 2014). DWLS estimation generally produces accurate results when analyzing medium and large single-level models.

Most MSEM simulation studies do not evaluate the impact of model size. Most MSEM simulation studies regardless of estimation method use small models: 1-2 factors per level with 3 indicators per factor (e.g., Asparouhov & Muthén, 2007; Depaoli & Clifton, 2015; Holtmann et al., 2016; Hox et al., 2012; Meuleman & Billiet, 2009; Ryu & West, 2009; Yuan & Bentler, 2007). However, MSEM applications often use medium to large models that estimate at least 4 to 7 factors at each level (Kim et al., 2016). Evaluating the impact of varying model size when using DWLS to estimate MSEM is an important need for future research. Future MSEM research could evaluate DWLS performance when modeling 2 (small model) and 4 factors (large model) at each level. Single-level SEM simulations generally suggest that DWLS likely will perform well even when estimating relatively large MSEMs.

Model misspecification. Model misspecification indicates that some aspect of the model is incorrect. Model misspecification is important to evaluate because SEM applications rarely know the "true" model. The model being fit in application generally is wrong to some extent. This area of research generally evaluates how misspecification affects fit indices. Ideally, the fit indices would indicate misfit given a wrong model is being fit. The specific aspect(s) of the model being misspecified varies widely. Sometimes it can be a nonzero parameter that is fixed to zero in the model. Misspecification also can be represented using collapsed factors. If the true model is 2 factors, the misspecified model might fit 1 factor where all items load onto the same factor.

Single-level DWLS evaluations with misspecified models are mixed. DWLS yielded fairly accurate parameter estimates regardless of degree of misspecifcation (Bandalos, 2014). Power to detect misspecified factor loadings was extremely low for highly asymmetric data and small samples (N = 100; Bandalos, 2014). Compared to ML, DWLS yielded slightly less accurate standard errors but slightly better Type I error rates and power to detect omitted paths (Lei, 2009). DWLS generally produced accurate fit indices when analyzing misspecified models (Yu & Muthén, 2002). This accuracy decreased as data asymmetry increased and sample size decreased.

Little MSEM research has evaluated misspecification's impact on DWLS. MSEM simulation studies that assess the effect of misspecification generally use ML estimation (Boulton, 2011; Hsu et al. 2015, 2017; Ryu & West, 2009; Yuan & Bentler, 2007). The effect of model misspecification when using DWLS to estimate MSEMS is an important area for future research. DWLS likely will perform poorly (e.g., low power to detect misfit) when analyzing misspecified models with few groups and severe asymmetry.

Level-specific fit indices. Little MSEM research has studied level-specific fit indices when using DWLS estimation. However, single-level SEM evaluations found differing performance of fit indices for DWLS and ML (Yu & Muthén, 2002). DWLS fit indices generally were slightly underpowered compared to ML-based fit indices (Yu & Muthén, 2002). Moreover, level-specific fit indices may be underpowered to detect Level-2 misfit when ICCs are low (Hsu et al., 2017). Thus, level-specific fit indices should be studied when using DWLS estimation.

Asymmetric data. Data distribution is important because it can affect the accuracy and power of estimators including DWLS (Bandalos, 2014; Lei, 2009). Categorical variables cannot be normally distributed. Thus, normal and non-normal data may be described as symmetric and asymmetric, respectively (*cf* Bandalos, 2014). Symmetry indicates approximate normally distributed data. Asymmetry indicates skewed and/or kurtotic data (e.g. many extreme scores in the tails). Asymmetry is common when analyzing attitudinal measures, which may make socially desirable statements. Many people may respond similarly to socially desirable statements by selecting the extreme answers on the ends of the response scale (e.g., agreeing that studying is very important).

Single-level investigations of asymmetry's effect typically use slight to moderate asymmetry (Bandalos, 2014). With moderately or very skewed data, DWLS generally has low power to detect misfit (Bandalos, 2014; Lei, 2009; Yu & Muthén, 2002). DWLS parameter estimates sometimes are slightly biased when data are moderately asymmetric (Lei, 2009). Bias of standard errors was generally minor when data were moderately asymmetric and were very small samples (N = 100; Lei, 2009). Most DWLS fit indices had low power to detect misfit with moderately asymmetric data and N = 100 (Yu & Muthén, 2002). DWLS fit indices accurately identified misspecification with moderately asymmetric data and N = 250 (Yu & Muthén, 2002). When the model is correctly specified, DWLS yields high power, acceptable Type I error, and accurate fit indices even with moderately asymmetric data (Finney et al., 2016; Li, 2016).

Most MSEM research regardless of estimation method does not evaluate the effect of asymmetry/skew. DWLS studies usually analyze continuous multivariate normal

data or categorical data without explicitly describing or varying the degree of asymmetry. Single-level evaluations suggest that DWLS may perform poorly when analyzing severely asymmetric/skewed data (e.g., Bandalos, 2014). Thus, future MSEM research should evaluate DWLS performance with asymmetric data.

Literature Review Summary and Need for Current Study

Attitudinal self-report measures are common in psychology and higher education. Students may evaluate their teacher's effectiveness (e.g. Sessoms & Willse, 2019). These questionnaires present several scoring and modeling challenges. They often are ordered categorical data, multidimensional, and given in multilevel contexts. MSEM generally is appropriate for analyzing and modeling these responses. MSEM can accommodate multidimensionality and provide estimates of relationships among latent variables while correcting for measurement error. MSEM explicitly accounts for the nested data structure and generally produces more accurate parameter estimates than SEM.

Important considerations in MSEM are sample size (number of groups and group size), model fit, ICCs, and estimation method. Increasing the number of groups, group size, and ICCs generally increases model convergence and parameter estimation accuracy. The most appropriate estimation method depends on the data (e.g., number of response categories, degree of asymmetry) and can affect accuracy and convergence. Most MSEM simulations have used ML estimation.

DWLS estimation has not been explored thoroughly in MSEM. Single-level SEM simulations using DWLS generally are promising. MSEM simulations are limited in that they generally fit the "correct" model only, use small models (1-2 factors at each level),

and do not evaluate the effect of asymmetry. These simulation characteristics often do not match MSEM applications. MSEM applications often fit large models (at least 4-7 factors at each level) that are partially misspecified and analyze asymmetric data (Kim et al., 2016). These limitations hold for most MSEM simulations regardless of estimation method. MSEM evaluations of DWLS have not studied the accuracy of level-specific fit indices. Single-level SEM evaluations found differing performance of fit indices for DWLS and ML (Yu & Muthén, 2002). This simulation study will address these literature gaps by studying the effects of model misspecification, large models, and asymmetric data on convergence and level-specific fit indices when estimating MSEMs using DWLS.

Research Questions

This study had two research questions. First, what are the convergence rates when fitting relatively large MSEMs using DWLS? Generally, in single-level SEM, fitting large models using DWLS results in non-convergence more frequently than when fitting small models (Bandalos, 2014). Moreover, what are convergence rates of large MSEMs when using relatively low ICCs and few groups? Generally, when multilevel data contain low ICCs and few groups, convergence of relatively simple MSEMs is somewhat low (e.g., Depaoli & Clifton, 2015). Second, how will level-specific fit indices perform under varying degrees of model misspecification, data asymmetry, and ICCs when using DWLS? When using ML estimation, level-specific fit indices tend to identify misfit more accurately as ICCs and model misspecification increase (Boulton, 2011; Hsu et al., 2017). Single-level DWLS fit indices generally identify misfit less accurately as model size and data asymmetry increases (Bandalos, 2014; Yu & Muthén, 2002).

This study is focused on evaluating level-specific fit indices' performance in relation to established cut-off criteria for those indexes in single-level contexts. My intent is *not* to determine new cutoff values for level-specific fit indices. That line of inquiry can be evaluated in future research.

CHAPTER III

METHODS

Overview of Methods

A simulation study was conducted to address the previously discussed limitations of MSEM simulations. Model misspecification, data asymmetry, and model size were studied using DWLS estimation. Two model size conditions, two data asymmetry conditions, seven model misspecification conditions, two ICC conditions, and two sample size conditions were evaluated. This simulation design yielded 112 conditions (2 x 2 x 7 x 2 x 2). Ryu and West's (2009) level-specific fit evaluation approach requires estimating five models for each factor structure tested. These models are Level-1 independence/Level-2 saturated, Level-1 theoretical/Level-2 saturated, Level-1 saturated/Level-2 independence, Level-1 saturated/Level-2 theoretical, and Level-1 theoretical/Level-2 theoretical. Using Ryu and West's (2009) level-specific fit evaluation approach and fitting multiple factor structures resulted in 10 models in each condition (5 level-specific fit models x 2 factor structures). Each condition was replicated 100 times.

Simulation Conditions

The conditions were data model size, asymmetry, model misspecification, ICC, and number of groups. Data conditions were chosen to align with previous simulations and increase generalizability to MSEM applications. Table 1 summarizes the conditions and values chosen.

Table 1. Summary of Simulation Conditions for Dissertation

Condition Type	Number of Conditions	Specific categories/values chosen
Model Size	2	2 factors at each level 4 factors at each level
Data Asymmetry	2	Symmetric Severe asymmetry
Model Misspecification	7	Within only (Level-1 only) misspecification 1) 1 cross loading per factor fixed to 0 2) Pairs of correlated factors collapsed into 1 (e.g. 4 true factors but 2 factors fit) Between Level only (Level-2 only) misspecification 3) 1 cross loading per factor fixed to 0 4) Pairs of correlated factors collapsed into 1 Both within/between (Level-1 and 2) misspecification 5) 1 cross loading per factor fixed to 0 6) Pairs of factors collapsed into 1 7) No misspecification (correct model fit at both levels)
Sample size	2	L2 N = 50 groups, Group size = 25, L1 N = 1250 L2 N = 100 groups, Group size = 25, L1 N = 2500
ICC	2	.15, .30

Note. This design yielded 112 conditions. ICC = intraclass correlation coefficient.

Model size condition. Two MCFA conditions were generated and associated models estimated. The small MCFA condition contained two factors at each level (Figure 1). The large MCFA condition contained four factors at each level (Figure 2). Both MSEMs contained four indicators per factor, implying 8 observed variables in the small model condition and 16 observed variables in the large condition (there are cross-loadings). Most MSEM simulations fit small models (1-2 factors per level). Fitting 2 factors at each level enabled comparability with previous simulations. Fitting 4 factors at each level, for the large model, used a higher number of factors consistent with actual practice. The average number of factors for multilevel factor analysis applications ranged from 3.5 to 4 at each level, with large variability (Kim et al., 2016).

Table 2 provides all population parameter values (e.g. "true" Level-1 factor loadings) except category thresholds for both models. Table 3 provides category thresholds. Two sets of population values were used to create two ICC conditions (see Table 2). Population values to create the ICCs were based on the MSEM observed ICC equation (Muthén, 1991):

$$\rho_i = (\lambda_{Bi}^2 \times \Psi_B + \Theta_B) / [(\lambda_{Bi}^2 \times \Psi_B + \Theta_B) + (\lambda_{Wi}^2 \times \Psi_W + \Theta_W)], \qquad (22)$$

where for item *i,* λ_{Bi}^2 is the squared between-level factor loading, Ψ_B is the between-level factor variance, Θ_B is the between-level error variance, λ_{Wi}^2 is the squared within level factor loading, Ψ_W is the within-level factor variance, and Θ_W is the within-level error variance.

All models had five types of parameters: factor variances, factor covariances, factor loadings, item thresholds, and error variances. Item thresholds and error variances were modeled at Level-2 only. Factor variances, factor covariances, and factor loadings were modeled separately for each level. All population parameters varied across ICC conditions except for Level-1 factor variances, Level-1 error variances, and Level-1 factor covariances. All Level-1 factor variances were fixed to 1 and Level-1 factor covariances were fixed to 0.5 to achieve factor correlations of .5. Given categorical data, Mplus automatically fixed all Level-1 error variances to 1 to identify the model (Asparouhov & Muthén, 2007). For a given level and ICC condition, factor variances and covariances had the same true value across all factors. Thresholds were modeled at Level-2 and had the same value across all primary loadings. Thresholds were higher for secondary items than primary items to reflect more variance in those items (see Table 3). For a given level and ICC condition, thresholds were the same across secondary items. Further explanation of thresholds in general, their population values, and corresponding rationale is provided in the Data Asymmetry section. For a given level and ICC condition, primary factor loadings, secondary factor loadings, and Level-2 error variances were the same for all items and factors. Giving a population parameter the same value across items and/or factors is relatively common in MSEM simulations (e.g. Boulton, 2011; Clifton & Depaoli, 2017).

Table 2. Population Parameter Values for Data Generation

ICC	λ_{Wp}	λ_{Ws}	λ_{Bp}	λ_{Bs}	θ_B	θ_W	ψ_B	ψ_W	ϕ_B	ϕ_W
.15	.9220	0.461	.3873	0.19365	.15	1	1.1766	1	0.42495	0.5
.30	.8367	0.41835	.5477	0.27385	.30	1	1.4289	1	0.34992	0.5

Note. ICC = intraclass correlation coefficient; λ_{Wp} = Within-group primary factor loadings; λ_{Ws} = Within-group secondary factor loadings for items that load on to a second factor (i.e. cross-loadings); λ_{Bp} = between-group primary factor loadings; λ_{Bs} = Between-group secondary factor loadings for items that load on to a second factor (i.e., cross loadings); θ_B = between-group residual variances; ψ_B = between-level factor variance; ψ_W = within-level factor variance; ϕ_B = between-level factor covariance; ϕ_W = within-level factor covariance. For categorical variables, Mplus automatically fixes the within-group residual variances θ_W to 1 to identify the model (Asparahouv & Muthén, 2007).

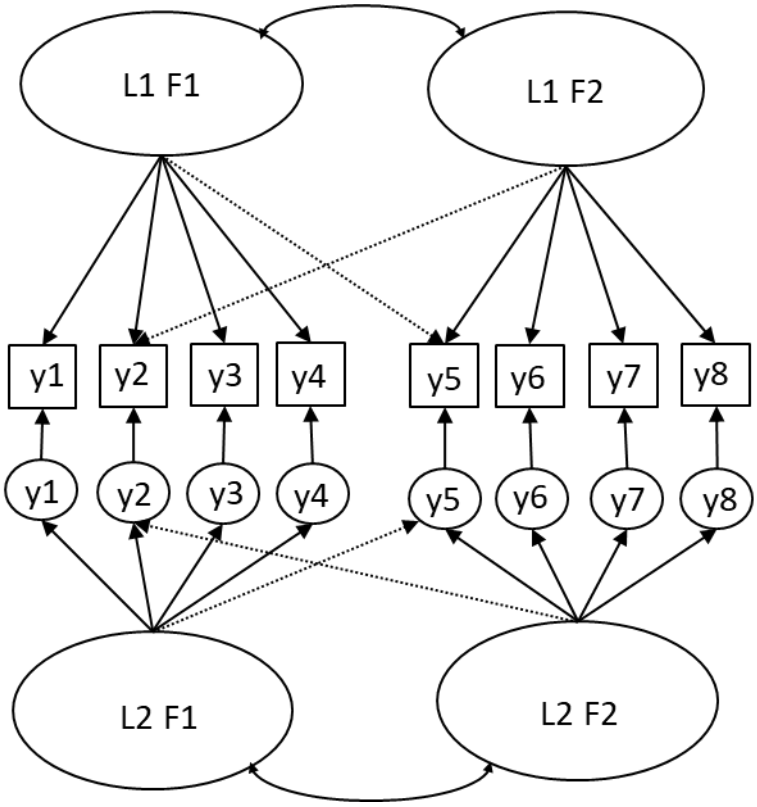

Figure 1. Population Small Model that Models 2 Factors at Each Level.
Note. Solid arrows indicate primary loadings of .7. Dotted arrows represent secondary loadings of .3.

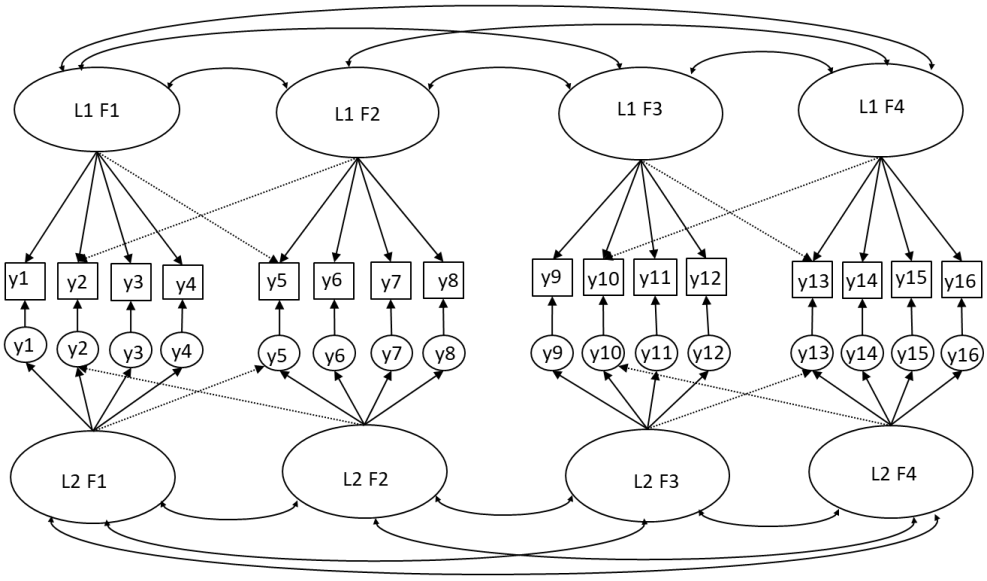

Figure 2. Population Large Model that Models 4 Factors at Each Level.
Note. Solid arrows represent primary loadings. Dotted arrows represent secondary loadings.

Data asymmetry condition. Two asymmetry conditions were studied: approximate symmetry and severe asymmetry. Approximate symmetry refers to data that is (roughly) normally distributed. Severely asymmetry describes data that is non-normal and heavily skewed in its distribution (e.g. many extreme scores). "Symmetry" and "asymmetry" were used instead of "normality" and "non-normality" because categorical data cannot be normally distributed. "Symmetry" and "asymmetry" are used often in the literature (*cf* Bandalos, 2014). In theory, DWLS is well-equipped to handle asymmetric data because it does not assume data normality like ML. However, DWLS generally performs poorly (e.g. low power to detect misfit) when analyzing severely asymmetric single-level data (Bandalos, 2014; Lei, 2009; Yu & Muthén, 2002). DWLS generally yields slight biased standard errors and parameter estimates when analyzing moderately asymmetric single-level data (Bandalos, 2014; Lei, 2009; Yu & Muthén, 2002).

Data were generated from a multivariate normal distribution. Data were discretized into four response categories. Once discretized, even the "normal" data condition is no longer formally "normal". The approximately symmetric condition had proportions of respondents corresponding to .15 (category 1), .35 (category 2), .35 (category 3), and .15 (category 4). The severely asymmetric condition had equally low proportions of responses in categories 1-3 and a large proportion of responses in category 4. These proportions correspond to values of .10, .10, .10, and .70. These proportions were created by specifying threshold values in Mplus. Thresholds indicate the z-score corresponding to the proportion of respondents who chose that category or below. The second threshold would indicate the z score corresponding to the proportion of examinees

who selected category 2 or 1. In the approximately symmetric condition, 15% of examinees chose category 1 and 35% chose category 2. Threshold 2 for the approximately symmetric condition would be the z-score corresponding to 50% of the respondents (.15 + .35). Table 3 provides all thresholds.

However, using z scores to determine population thresholds is not accurate when estimating multilevel models. This inaccuracy occurs because the z scores assume a variance of 1. However, in multilevel contexts, items contain variance at multiple levels, and likely have total variances exceeding 1. This problem was addressed by estimating the total variance of an item in a MSEM context, then multiplying the square root of this new variance by the old threshold that assumes a variance of 1. This item variance estimation required separate equations for items that did not cross-load and items that did.

The total variance of an item that did not cross-load was estimated using

$$\sigma_T^2 = \lambda_B^2 \psi_B + \lambda_W^2 \psi_W + \theta_B + \theta_W , \tag{23}$$

where for a given item, σ_T^2 is the total variance, λ_B^2 is the squared between-level factor loading, ψ_B is the between-level factor variance, λ_W^2 is the squared within-level factor loading, ψ_W is the within-level factor variance, θ_B is the between-level error variance, and θ_W is the within-level error variance. Table 2 gives specific values for each element.

The total variance of an item that did cross-load was estimated using

$$\sigma_T^2 = \Lambda_B' \Sigma_B \Lambda_B + \Lambda_W' \Sigma_W \Lambda_W + \theta_B + \theta_W , \tag{24}$$

where σ_T^2 is the total variance, Σ_B is the between covariance matrix of the two factors that item *i* measures, Λ_B is a vector of between-level factor loadings that contain the primary loading and secondary loading, Σ_W is the within-level covariance matrix of the two factors that item *i* measures, Λ_W is a vector of within-level factor loadings that contain the primary loading and secondary loading for item *i*, θ_B is the between-level error variance for item *i*, and θ_W is the within-level error variance for item *i*. The specific values for each element are provided in Table 2.

Separate sets of thresholds were used for the ICC = .15 and ICC = .30 conditions because ICC = .30 had more total variance than ICC = .15. Equation 23 and Table 2 for non-crossloading items yielded item total variance estimates of 2.71658 for ICC = .15 and 2.4287 for ICC = .30 items. Equation 24 and Table 2 for cross-loading items yielded item total variance estimates of 3.074642 for ICC = .15 and 3.416315 for ICC = .30 items. The square root of these values was multiplied for the corresponding "old" threshold in Table 3 to create the new thresholds.

Table 3. Thresholds Used to Generate Data Asymmetry Conditions

Threshold	Symmetric (old z-score)		Asymmetric (old z-score)	
τ_1	-1.0364		-1.2816	
τ_2	0.00		-0.8416	
τ_3	1.0364		-0.5244	
		ICC = .15		
New Threshold	Symmetric Not crossloading	Symmetric crossloading	Asymmetric Not crossloading	Asymmetric Crossloading
τ_1	-1.52902	-1.81729	-1.89077	-2.24724
τ_2	0.00	0.00	-1.24163	-1.47572
τ_3	1.52902	1.81729	-0.77366	-0.91952
		ICC = .30		
New Threshold	Symmetric Not crossloading	Symmetric crossloading	Asymmetric Not crossloading	Asymmetric Crossloading
τ_1	-1.61516	-1.91561	-1.99728	-2.36882
τ_2	0.00	0.00	-1.31157	-1.55555
τ_3	1.61516	1.91561	-0.81724	-0.96926

Model misspecification. There were seven misspecification conditions for both the small and large models (see Table 1). Collapsed factors or misspecified factor loadings will be studied for Level-1 only, Level-2 only, and both levels. The correct model (correctly specified at both levels) was fit to provide a baseline evaluation. Collapsed factors and/or misspecified factor loadings are typical in MSEM simulation studies that evaluate level-specific fit indices (e.g., Boulton, 2011; Hsu et al., 2015; 2017; Ryu & West, 2009). Collapsed factors are much more common than misspecified factor loadings in MSEM simulations (*cf* Boulton, 2011; Hsu et al., 2017; Ryu & West, 2009).

Collapsed factors usually are created by incorrectly collapsing pairs of correlated factors. Consider a MCFA that contains 2 factors at each level. Fitting all items to one factor at each level would result in incorrectly collapsed factors. Misspecified factor loadings are much more common in single-level SEM simulations (e.g., Bandalos, 2014; Lei, 2009) than in MSEM simulations (e.g., Hsu et al., 2015). Misspecified factor loadings typically are created by fixing a non-zero loading to zero. These misspecified factor loadings usually are cross-loadings. Cross-loadings indicate that an item measures multiple factors.

There were 7 misspecification conditions. Table 1 describes each misspecification condition. Figures 3 through 14 provide the misspecified model for Conditions 1-6 for the two model size conditions. Figures 3, 5, 7, 9, 11, and 13 depict Conditions 1 through 6, respectively, for the small model. Figures 4, 6, 8, 10, 12, and 14 depict Conditions 1 through 6, respectively, for the large model. Condition 7 will fit the true model which was shown in Figures 1 (small model) and 2 (large model).

Each level had two misspecification conditions: cross-loadings fixed to zero and collapsed factors. Conditions 1, 3, and 5 fixed cross-loadings to zero. Condition 1 fixed all Level-1 cross-loadings to zero (Figures 3 and 4). Condition 3 fixed all Level-2 cross-loadings to zero (Figures 7 and 8). Condition 5 combined Conditions 1 and 3 by fixing all Level-1 and Level-2 cross-loadings to 0 (Figures 11 and 12). Because the true model is identical at both levels, the number of misspecified loadings for Conditions 1 and 3 were identical. Condition 5 doubled the number of misspecified loadings in Conditions 1 and 3. Figures 3-8 also note the number of misspecified cross-loadings for each misspecification condition for each model size condition.

Conditions 2, 4, and 6 fit collapsed factors at one or both levels. Condition 2 collapsed Level-1 factors and correctly modeled the Level-2 factor structure (Figures 5 and 6). Condition 4 collapsed Level-2 factors and correctly modeled the Level-1 factor structure (Figures 9 and 10). Condition 6 collapsed both Level-1 and Level-2 factors (Figures 13 and 14).

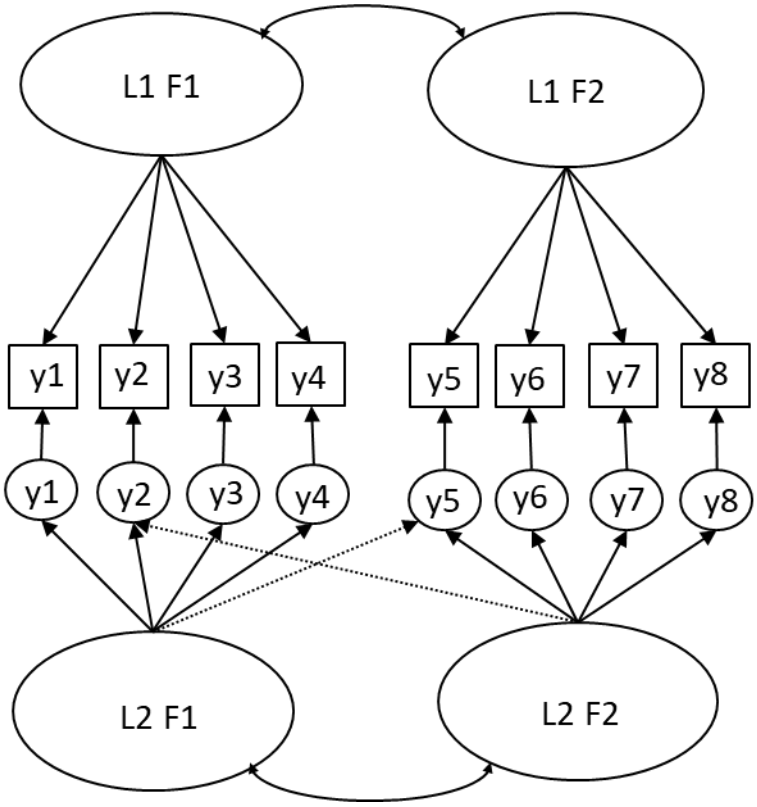

Figure 3. Misspecification Condition 1 for the Small Model.
Note. The misspecified model incorrectly fixes 2 cross-loadings at Level-1 to 0. The Level-2 model is correctly specified.

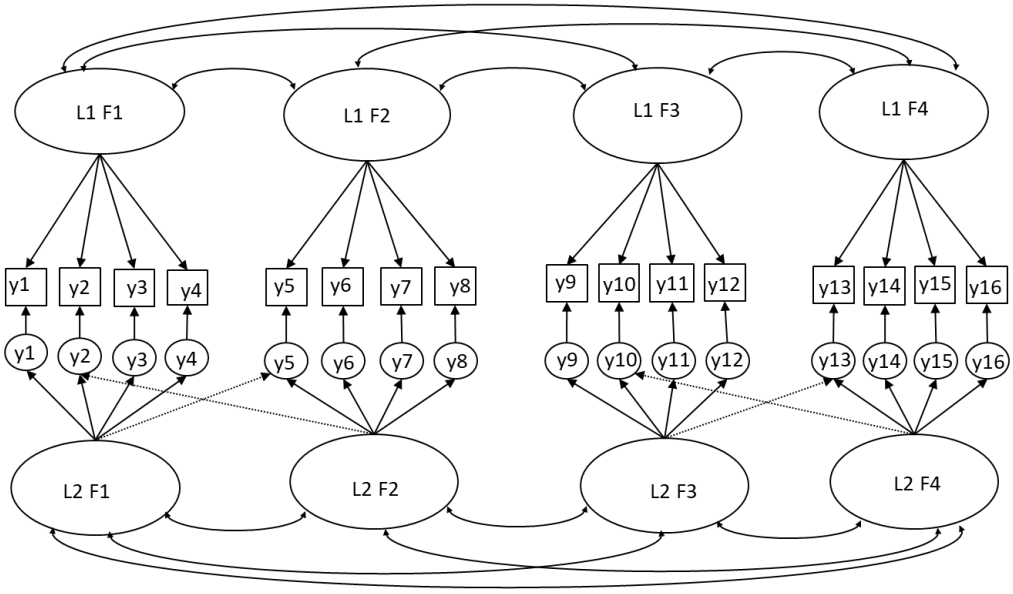

Figure 4. Misspecification Condition 1 for the Large Model.
Note. The misspecified model incorrect fixes 4 Level-1 cross-loadings to zero. The Level-2 model is correctly specified.

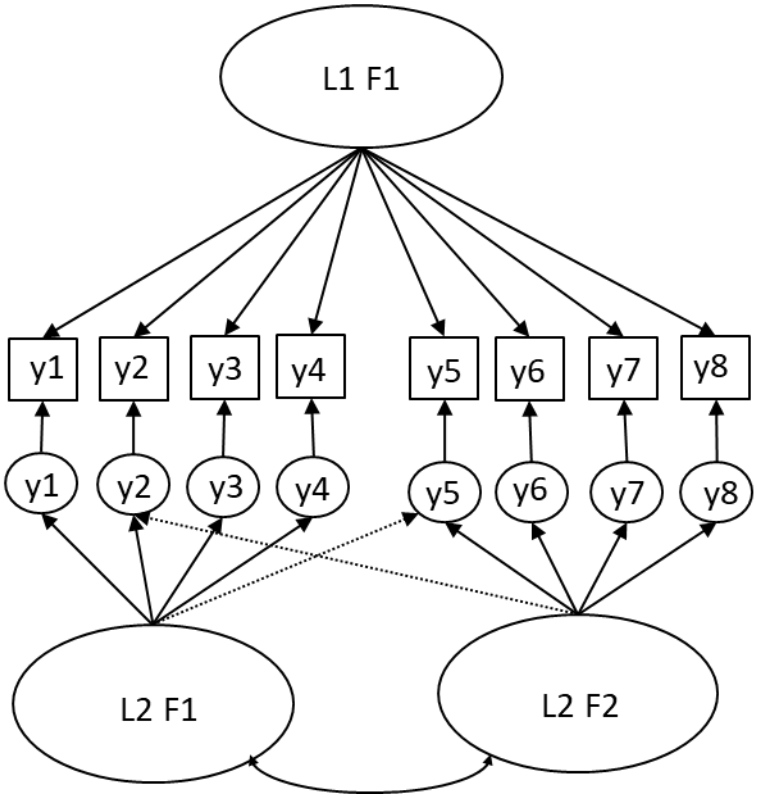

Figure 5. Misspecification Condition 2 for the Small Model.
Note. The misspecified model incorrectly fits 1 factor at Level-1. The Level-2 model is correctly specified.

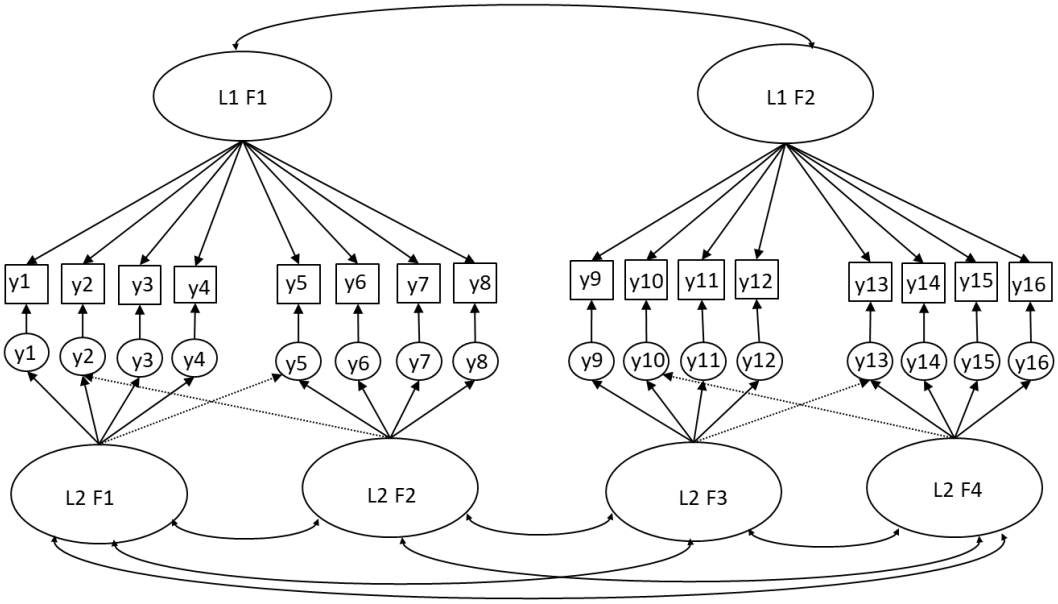

Figure 6. Misspecification Condition 2 for the Large Model.
Note. The model incorrectly fits 2 factors at Level-1. The Level-2 model is correctly specified.

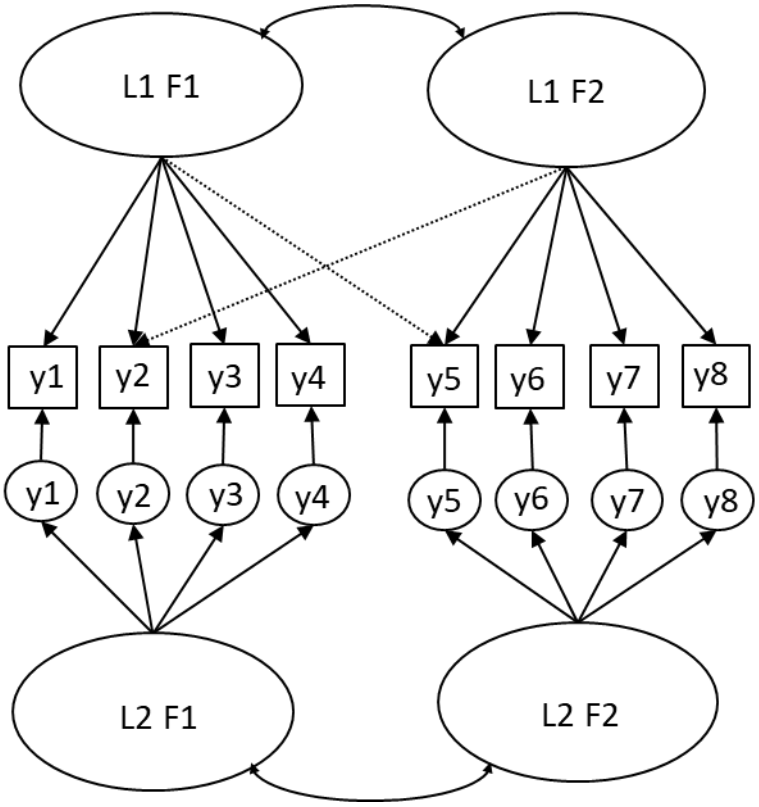

Figure 7. Misspecification Condition 3 for the Small Model.
Note. The model incorrectly fixes two Level-2 cross-loadings to zero. The Level-1 model is correctly specified.

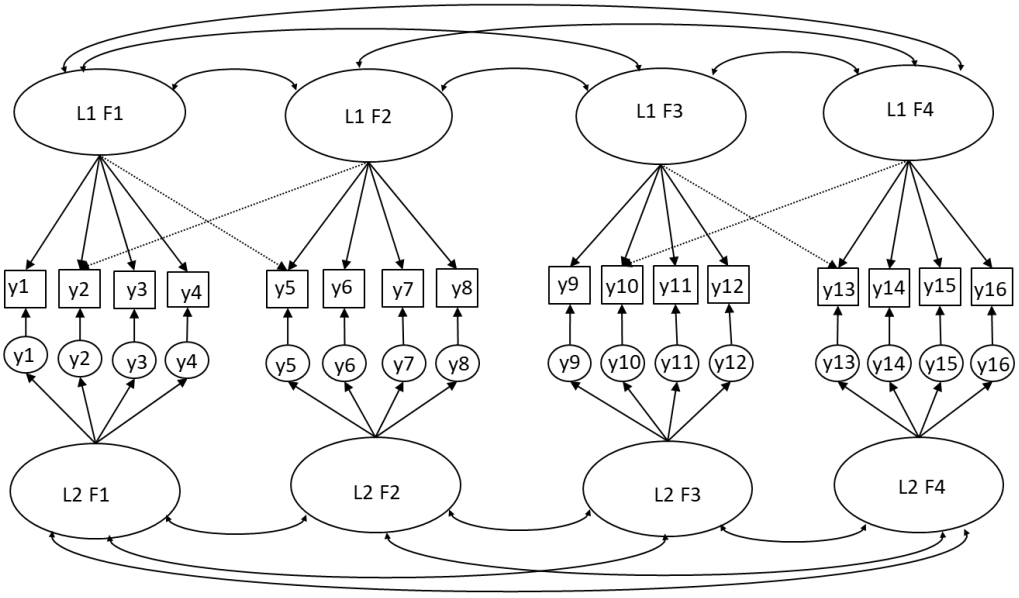

Figure 8. Misspecification Condition 3 for the Large Model.
Note. The misspecified model incorrectly fixes 4 Level-2 cross-loadings to zero. The Level-1 model is correctly specified.

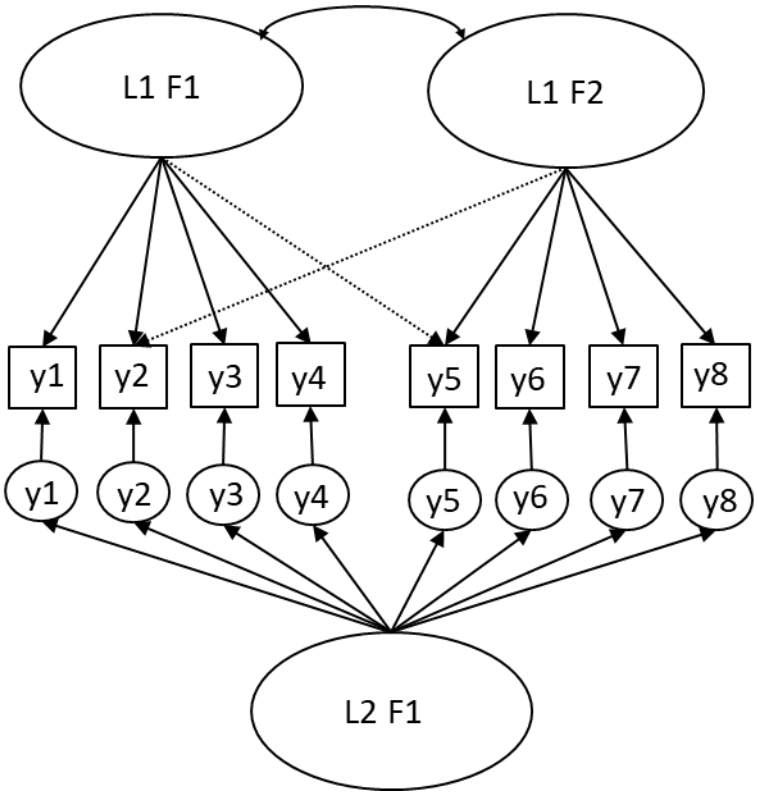

Figure 9. Misspecification Condition 4 for the Small Model.
Note. The misspecified model incorrectly models 1 factor at Level-2. The Level-1 model is correctly specified.

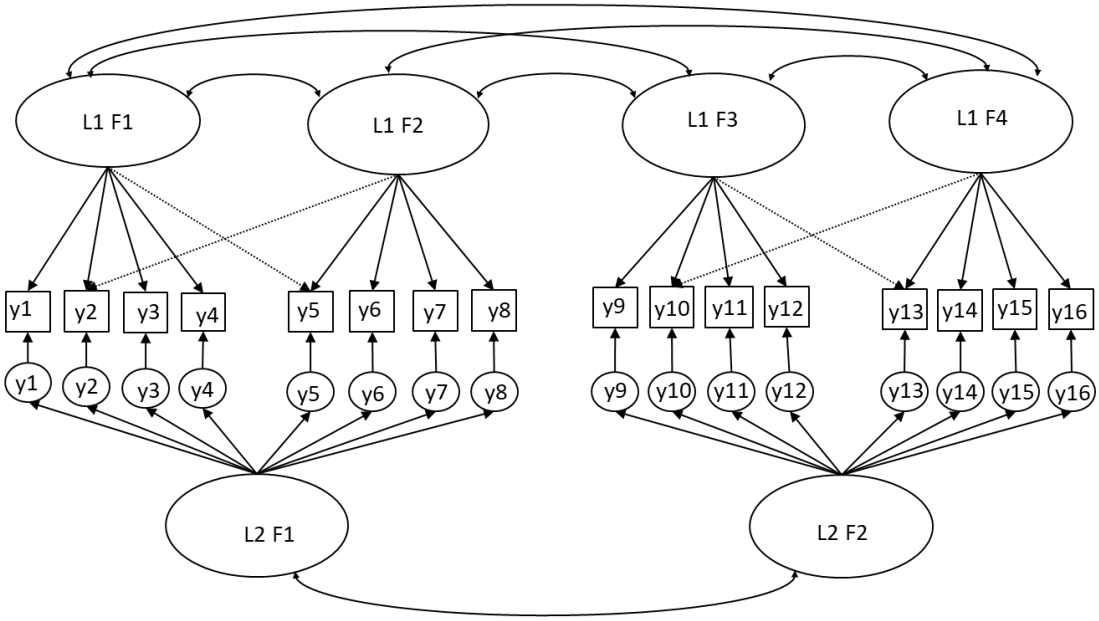

Figure 10. Misspecification Condition 4 for the Large Model.
Note. The misspecified model incorrectly fits 2 factors at Level-2. The Level-1 model is correctly specified.

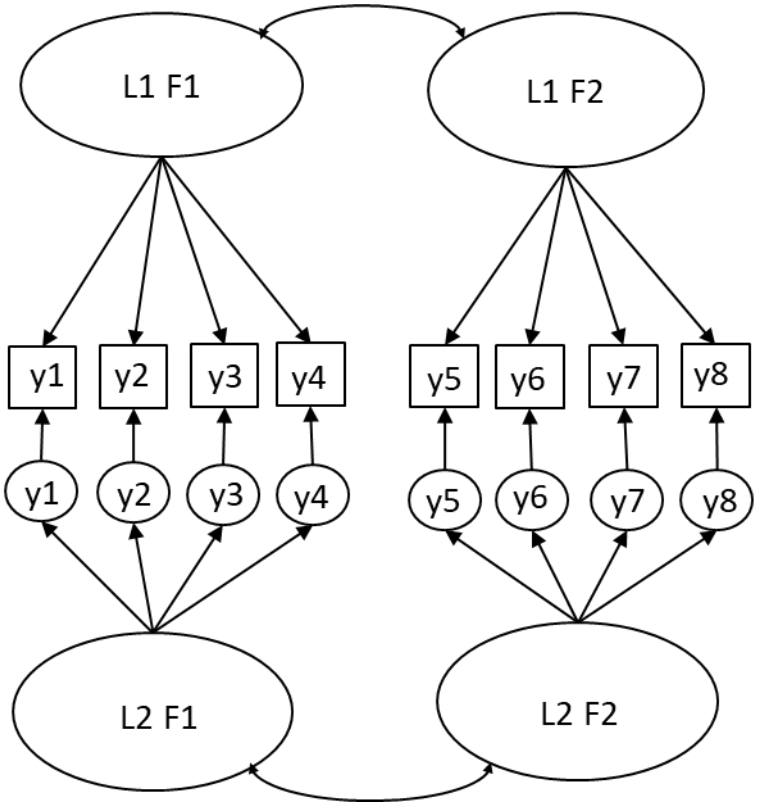

Figure 11. Misspecification Condition 5 for the Small Model.
Note. The model incorrectly fixes 2 Level-1 cross-loadings and 2 Level-2 cross-loadings to zero.

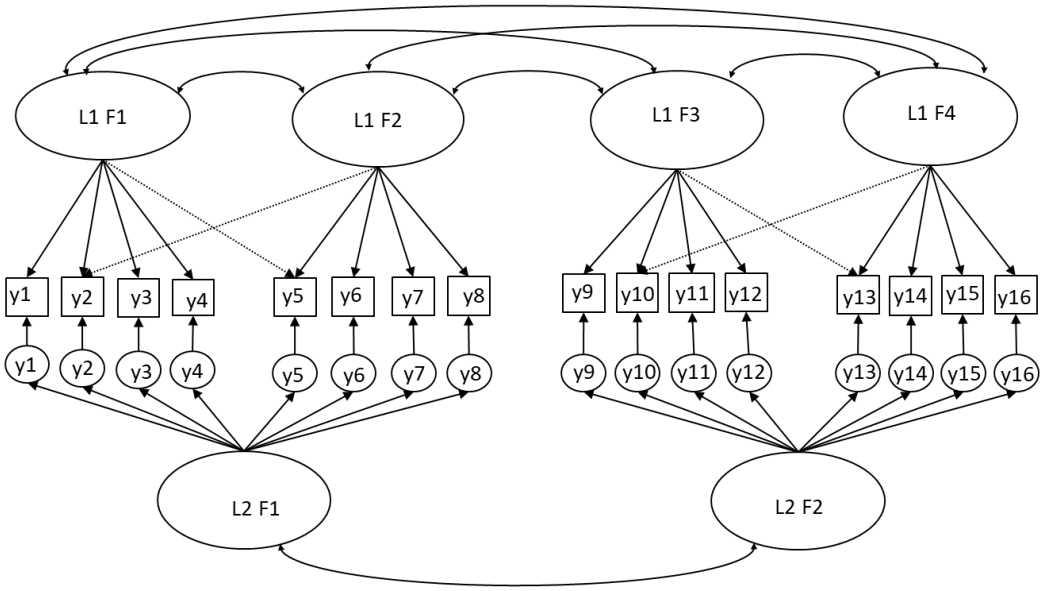

Figure 12. Misspecification Condition 5 for the Large Model.
Note. The model incorrectly fixes four Level-1 cross-loadings and four Level-2 cross-loadings to zero.

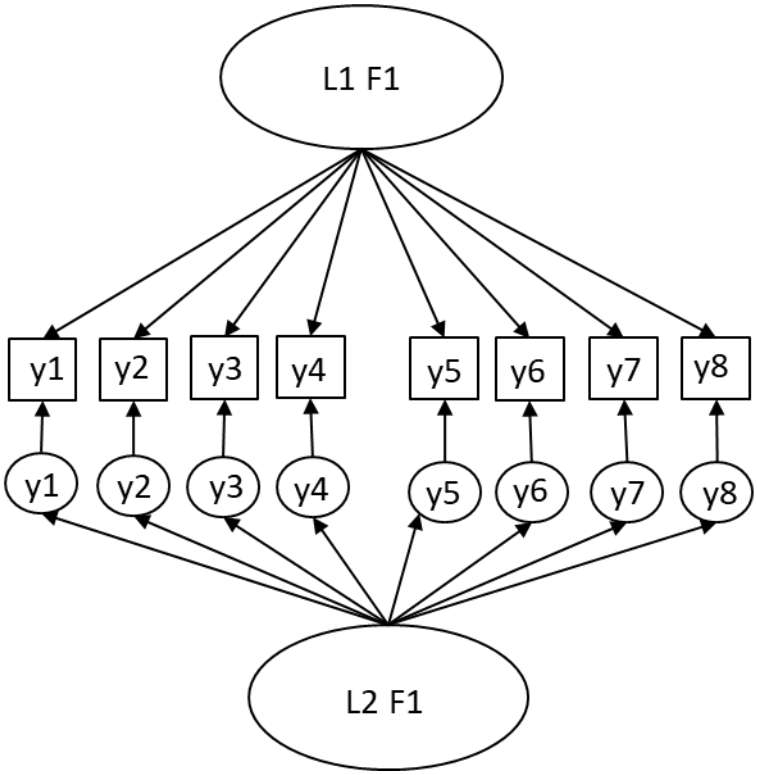

Figure 13. Misspecification Condition 6 for the Small Model.
Note. The misspecified model incorrectly fits 1 Level-1 factor and 1 Level-2 factor.

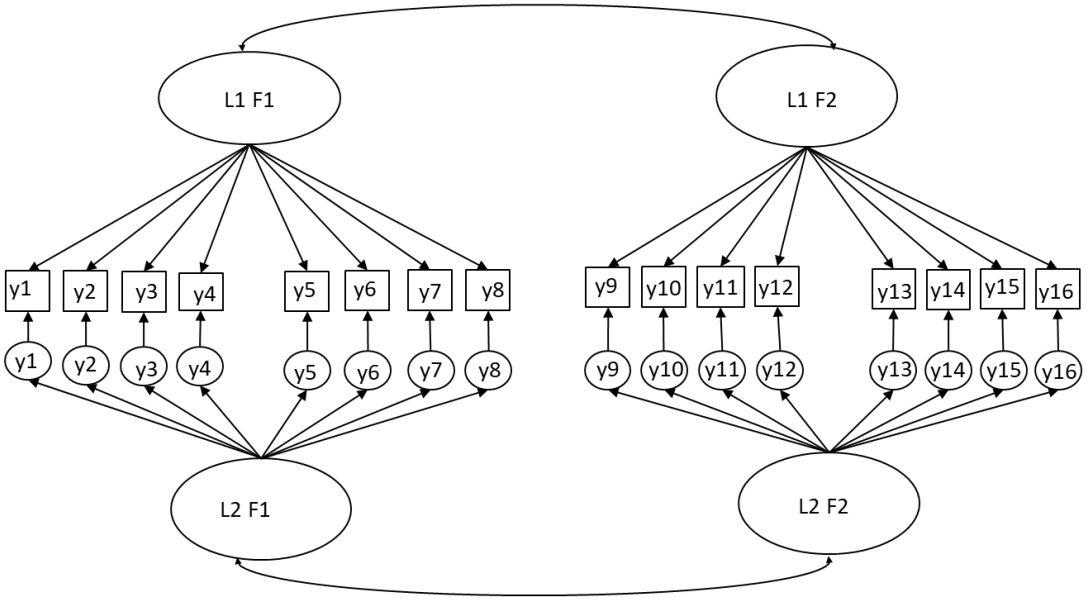

Figure 14. Misspecification Condition 6 for the Large Model.
Note. The misspecified model incorrectly fits 2 Level-1 factors and 2 Level-2 factors.

ICC conditions. There were two ICC conditions: small (ICC = .15) and medium (ICC = .30). Observed ICCs were used given their more frequent use in simulations and applications. These values are supported by MSEM applications and simulations. A literature review of multilevel factor analysis applications found that the average minimum and maximum observed ICCs were .13 and .34 (Kim et al. 2016). Observed ICCs > .30 generally are rare in application (Hox, 2010; Ludtke et al., 2008). MSEM simulation studies typically use ICC conditions that range from .05 to .30 or .40 (e.g., Boulton, 2011; Depaoli & Clifton, 2015; Guenole, 2016; Ludtke et al., 2008; Ludtke et al., 2011; Preacher et al., 2011). ICC of .15 was chosen as the small ICC value because ICCs < .10 can cause non-convergence (Boulton, 2011; Clifton & Depaoli 2017; Depaoli & Clifton, 2015). Moreover, large models like those being fit in my study generally require larger ICC values to converge (Preacher et al., 2016). Thus, ICCs of .15 were chosen as baseline values.

Sample size. There were two sample size conditions. The number of groups was varied, but group size was held constant at 25. Number of groups was 50 (small) and 100 (large). A literature review of multilevel factor analysis applications found that the median number of groups was 83 (Kim et al., 2016). The average number of groups was 177. MSEM simulation studies typically use 50, 100, and 200 groups (e.g., Depaoli & Clifton, 2015; Holtmann et al., 2016; Hox et al., 2010). Groups of 50 and 100 were chosen to align with typical MSEM applications and simulations.

Group size of 25 is typical of MSEM applications and simulations. A literature review of multilevel factor analysis applications found that the median group size was 18

and average group size was 26 (Kim et al., 2016). Group sizes of 20 or 25 are common in MSEM simulations (e.g., Boulton, 2011; Depaoli & Clifton, 2015; Guenole, 2016; Hox et al., 2010; Preacher et al., 2011). Groups of 20 or 25 represent moderate to large group sizes in MSEM simulations. Generally, a range of group sizes is used in MSEM simulations. I did not vary group size because it was not the study's focus. However, I wanted to use group sizes that are typical of MSEM applications and simulations.

The total Level-1 sample size was computed by multiplying the number of groups by the group size. The Level-1 sample size for 50 groups condition was 1,250 (50 groups x 25 group members). The Level-1 sample size for 100 groups condition was 2,500 (100 groups x 25 group members). These values are somewhat small for MSEM applications. A literature review of multilevel factor analysis applications found that the median Level-1 sample size was 1,500 and the mean Level-1 sample size was 4,500 (Kim et al., 2016).

Evaluation of Results and Analytic Approach

The primary dependent variables were the level-specific fit indices described in Chapter 2. Level-specific CFI, TLI, RMSEA, and SRMR were evaluated given their predominant usage in MSEM simulations (e.g., Hsu et al., 2017). General guidelines of CFI $\geq$ .95, TLI $\geq$ .95, RMSEA $\leq$ .06, and SRMR $\leq$ .08 were used to guide model fit evaluation (Hu & Bentler, 1999). Convergence rates also were inspected and reported. Convergence rates were the proportion of replications for each condition that successfully converged (out of 100 replications).

Mplus 8.1 was used to generate and analyze datasets. Each replication result (dataset, parameter estimates, standard errors, and fit indices) for each condition was saved. Mplus does not automatically compute level-specific fit indices (except for level-specific SRMR). Level-specific fit indices were computed manually using R. The level-specific fit indices were computed using five models for each condition. R 3.7 was used to summarize results and create tables and figures.

CHAPTER IV

RESULTS

This simulation study had two primary research questions. First, what were the convergence rates? Second, how would level-specific fit indices perform under conditions of model misspecification, data asymmetry (non-normality), number of groups, ICCs, and model size? Hu and Bentler's (1999) cut-off values of CFI $\geq .95$, TLI $\geq .95$, RMSEA $\leq .06$, and SRMR $\leq .08$ often are used to indicate good model fit. These values were applied to the level-specific fit indices to estimate Type I and II error. Type I error indicates that the fit indices suggest misfit despite fitting the correct model. Type II error indicates that the fit indices suggest good fit despite fitting an incorrect model.

Model Convergence Results

Convergence was excellent. Appendices A through D provide all convergence rates. Small model conditions had 95% to 100% convergence. All small model data conditions had perfect convergence except with few groups and small ICCs. Large model conditions had 93% to 100% convergence. The lowest convergence rates for large models occurred with few groups and small ICCs.

Model Fit Index Performance

Level-1, Level-2, and aggregate fit indices' performance was evaluated. Level-1 and Level-2 fit indices were defined previously in Chapter 2. Aggregate fit indices are fit indices that combine (mis)fit across both levels (i.e., not level-specific). These fit indices

are computed automatically when estimating MCFA. I inspected the aggregate fit indices provided when fitting the full MCFA (theoretical factor model at both levels). Model fit index performance was defined using the proportion of replications that indicated misfit using Hu and Bentler's (1999) suggested good fit values. Consider CFI and its suggested good fit value of ≥ .95. Models with CFI values outside this desired range of ≥ .95 would be considered poor fitting models and thus rejected. If 75 of 100 replications yielded CFI values ≤ .95 (i.e., 75 replications indicated poor model fit), then the rejection rate would be .75 (75/100). Throughout this chapter I use both percentages (e.g. rejected 75% of replications) and proportions (e.g. a .75 rejection rate).

A chapter outline follows. I first will describe Level-1 fit performance, then Level-2 fit performance, and finally aggregate fit index performance. I will conclude by comparing performances of Level-1, Level-2, and aggregate fit indices. Within a given section (e.g., Level-1 fit index performance), I first will outline general trends, then describe rejection rates for data asymmetry, model size, number of groups, and ICCs. For a given condition (e.g., asymmetry), I will describe model rejection results for cross-loadings fixed to zero, collapsed factors, and the correct model.

Level-1 Fit Indices' Overall Performance

Table 4 provides small model rejection rates for Level-1 fit indices. Table 5 provides large model rejection rates for Level-1 fit indices.

Table 4. Rejection Rates of Level-1 Fit Indices for Small Model Conditions

Fit Index	L1 Misp. Cross-loadings	L1 Collapsed Factors	Correct Model	Data
L1 CFI	0	0.98	0	Skewed data, 50 groups, ICC = .15
L1 TLI	0.15	1	0	
L1 RMSEA	0.05	1	0	
L1 SRMR	0	0.94	0	
L1 CFI	0	1	0	Skewed data, 100 groups, ICC = .15
L1 TLI	0.08	1	0	
L1 RMSEA	0.01	1	0	
L1 SRMR	0	0.97	0	
L1 CFI	0	0.82	0	Skewed data, 50 groups, ICC = .30
L1 TLI	0.07	0.98	0	
L1 RMSEA	0	0.65	0	
L1 SRMR	0	0.68	0	
L1 CFI	0	0.99	0	Skewed data, 100 groups, ICC = .30
L1 TLI	0.02	1	0	
L1 RMSEA	0	0.92	0	
L1 SRMR	0	0.67	0	
L1 CFI	0	1	0	Normal data, 50 groups, ICC = .15
L1 TLI	0.11	1	0	
L1 RMSEA	0.73	1	0	
L1 SRMR	0	0.92	0	
L1 CFI	0	1	0	Normal data, 100 groups, ICC = .15
L1 TLI	0.15	1	0	
L1 RMSEA	0.89	1	0	
L1 SRMR	0	1	0	
L1 CFI	0	0.94	0	Normal data, 50 groups, ICC = .30
L1 TLI	0.07	1	0	
L1 RMSEA	0.09	1	0	
L1 SRMR	0	0.62	0	
L1 CFI	0	1	0	Normal data, 100 groups, ICC = .30
L1 TLI	0.02	1	0	
L1 RMSEA	0.16	1	0	
L1 SRMR	0	0.54	0	

Note. Misp = misspecified.

Table 5. Rejection Rates of Level-1 Fit Indices for Large Model Conditions

Fit Index	L1 Misp. Cross-loadings	L1 Misp Factors	Correct Model	Data
L1 CFI	0	0.64	0	Skewed data, 50 groups, ICC = .15
L1 TLI	0	0.88	0	
L1 RMSEA	0	0	0	
L1 SRMR	0	0.04	0	
L1 CFI	0	0.82	0	Skewed data, 100 groups, ICC = .15
L1 TLI	0	0.97	0	
L1 RMSEA	0	0	0	
L1 SRMR	0	0	0	
L1 CFI	0	0.44	0	Skewed data, 50 groups, ICC = .30
L1 TLI	0	0.64	0	
L1 RMSEA	0	0	0	
L1 SRMR	0	0	0	
L1 CFI	0	0.52	0	Skewed data, 100 groups, ICC = .30
L1 TLI	0	0.85	0	
L1 RMSEA	0	0	0	
L1 SRMR	0	0	0	
L1 CFI	0	0.87	0	Normal data, 50 groups, ICC = .15
L1 TLI	0	0.99	0	
L1 RMSEA	0	0.6	0	
L1 SRMR	0	0	0	
L1 CFI	0	1	0	Normal data, 100 groups, ICC = .15
L1 TLI	0	1	0	
L1 RMSEA	0	0.97	0	
L1 SRMR	0	0	0	
L1 CFI	0	0.49	0	Normal data, 50 groups, ICC = .30
L1 TLI	0	0.82	0	
L1 RMSEA	0	0	0	
L1 SRMR	0	0	0	
L1 CFI	0	0.82	0	Normal data, 100 groups, ICC = .30
L1 TLI	0	1	0	
L1 RMSEA	0	0	0	
L1 SRMR	0	0	0	

Note. Misp = misspecified.

Rejection of Level-1 cross-loadings fixed to 0. Level-1 CFI and Level-1 SRMR never rejected Level-1 cross-loadings fixed to 0 regardless of data condition. Level-1 TLI and RMSEA never rejected Level-1 cross-loadings fixed to 0 for large models. For small models, Level-1 TLI rejected between 0% and 15% of Level-1 cross-loadings fixed to 0. Level-1 RMSEA's power to reject Level-1 cross-loadings fixed to 0 generally was zero or $\leq .10$ for most conditions. Level-1 RMSEA never rejected Level-1 cross-loadings fixed to 0 when fitting large models. Level-1 RMSEA rejected between 0% and 16% of small models with two exceptions. For small models, Level-1 RMSEA performed best with symmetric data and small ICCs. In these situations, Level-1 RMSEA rejection rates were .73 with few groups and .89 with many groups.

Rejection of Level-1 collapsed factors. Level-1 CFI rejection rates were much lower for large collapsed models (.44 to 1) than small collapsed models (.82 to 1). For large models, Level-1 CFI performed worst with severe asymmetry and large ICCs, yielding rejection rates of .44 to .52. For large models, Level-1 CFI performed best with symmetry and small ICCs, yielding rejection rates of .87 to 1. For small models, Level-1 CFI performed worst with severe asymmetry, few groups, and large ICCs (rejection rate of .82). Otherwise, Level-1 CFI rejection rates were .94 to 1.0 for large collapsed models.

Level-1 TLI rejection rates were much lower for large collapsed models (.64 to 1) than small collapsed models (.98 to 1). With large models, Level-1 TLI performed worst with severe asymmetry, few groups, and large ICCs, yielding a rejection rate of .64. Otherwise, Level-1 TLI rejected 85% to 100% of large models with collapsed Level-1 factors.

Level-1 RMSEA's power to detect Level-1 collapsed factors was zero for most large model conditions. For large models, Level-1 RMSEA performed best with symmetric data and low ICCs. In these situations, Level-1 RMSEA rejection rates were .60 with few groups and .97 with many groups. With small models, Level-1 RMSEA power was 1 in six of eight conditions. With small models, Level-1 RMSEA performed worst with asymmetric data, few groups, and large ICCs (rejection rate = .65). Otherwise, Level-1 RMSEA rejected between 92% and 100% of small collapsed models.

For Level-1 collapsed factors, Level-1 SRMR rejection rates were 0 to .04 for small models and .54 to 1 for large models. Level-1 SRMR performed best with small models and small ICCs, yielding rejection rates of .92 to 1. Otherwise, Level-1 SRMR rejection rates of small models with collapsed Level-1 factors ranged from .54 to .68.

Rejection of correct Level-1 models. Level-1 CFI, TLI, RMSEA and SRMR never rejected the correct model regardless of data condition.

Data Asymmetry's Impact on Level-1 Fit Indices

Before describing data asymmetry's impact on Level-1 fit indices' rejection rates, figures will be shown. These figures introduce the complex nature of level-specific fit index performance. Data asymmetry's impact generally also depended on the specific fit index, type of misspecification, model size, number of groups, and ICCs. Figures 15 through 18 contrast Level-1 fit indices' rejection rates with symmetric and severely asymmetric data for various combinations of conditions. Tables 4 and 5 provide rejection rates for Level-1 fit indices. Appendix E provides the category proportions for one of the asymmetry condition replications to illustrate the data's skewed distribution.

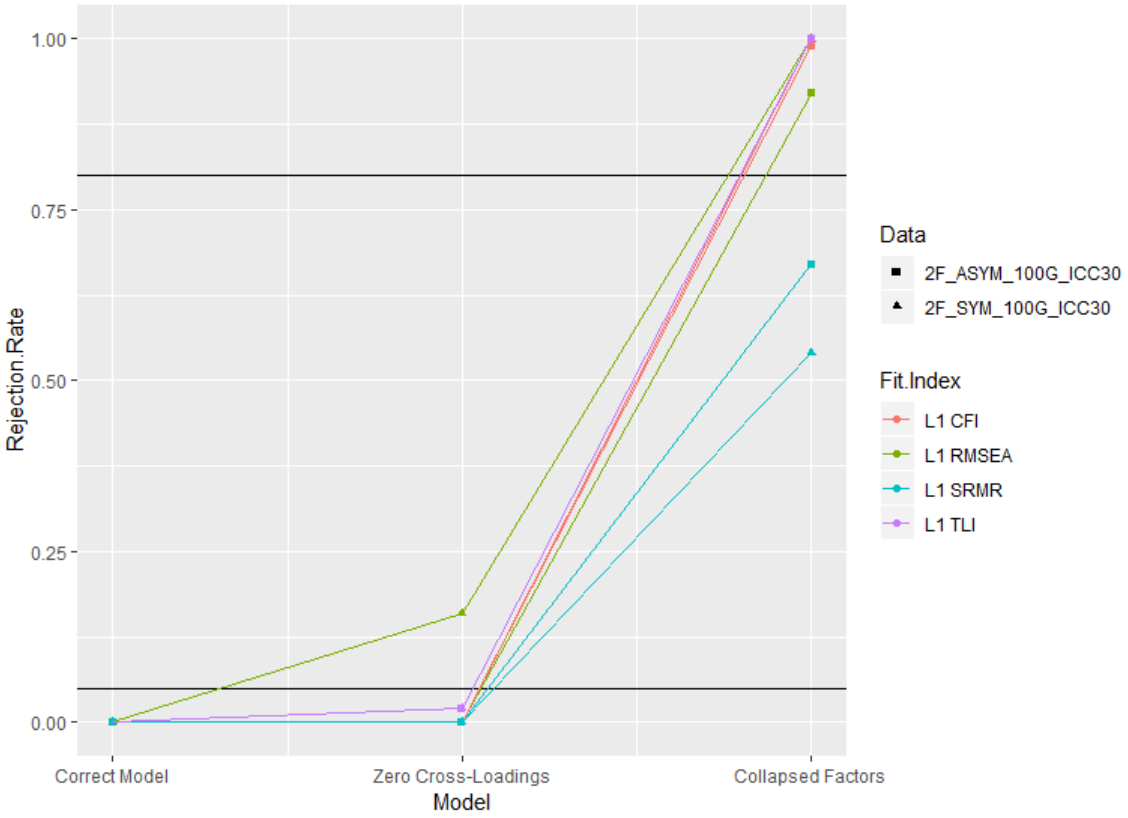

Figure 15. Impact of Data Asymmetry on Level-1 Fit Indices' Rejection Rates with Small Models, Many Groups, and Large ICCs.
Note. This figure has substantial overlap. All Level-1 fit indices never rejected correct models regardless of asymmetry. For zero cross-loadings, rejection rates were zero for Level-1 CFI, RMSEA, and SRMR with asymmetry and Level-1 CFI and SRMR with symmetry. For collapsed factors, rejection rates were 1 for Level-1 CFI, TLI, and RMSEA with symmetry and Level-1 TLI with asymmetry.

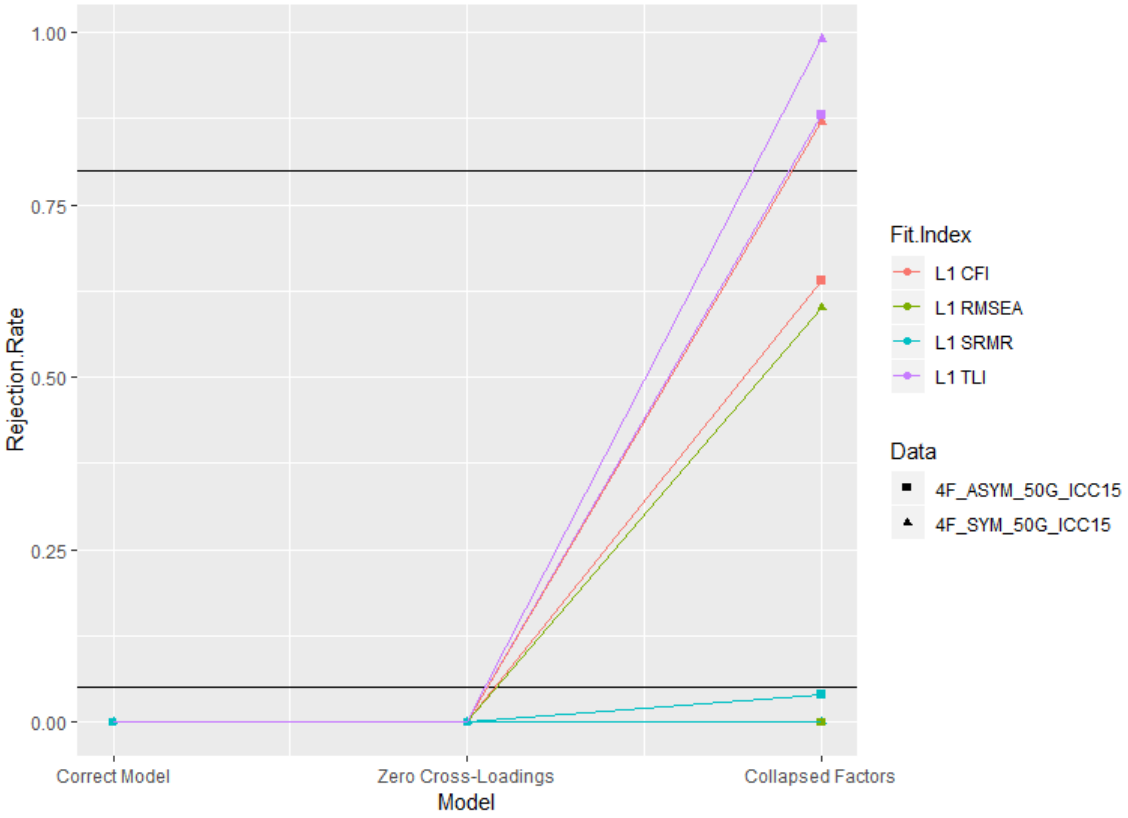

Figure 16. Impact of Data Asymmetry on Level-1 Fit Indices' Rejection Rates with Large Models, Few Groups, and Small ICCs.

Note. This figure has considerable overlap. All Level-1 fit indices' rejection rates were zero for correct models and zero cross-loadings regardless of symmetry. Collapsed factor rejection rates were zero for Level-1 SRMR with symmetry and Level-1 RMSEA with asymmetry.

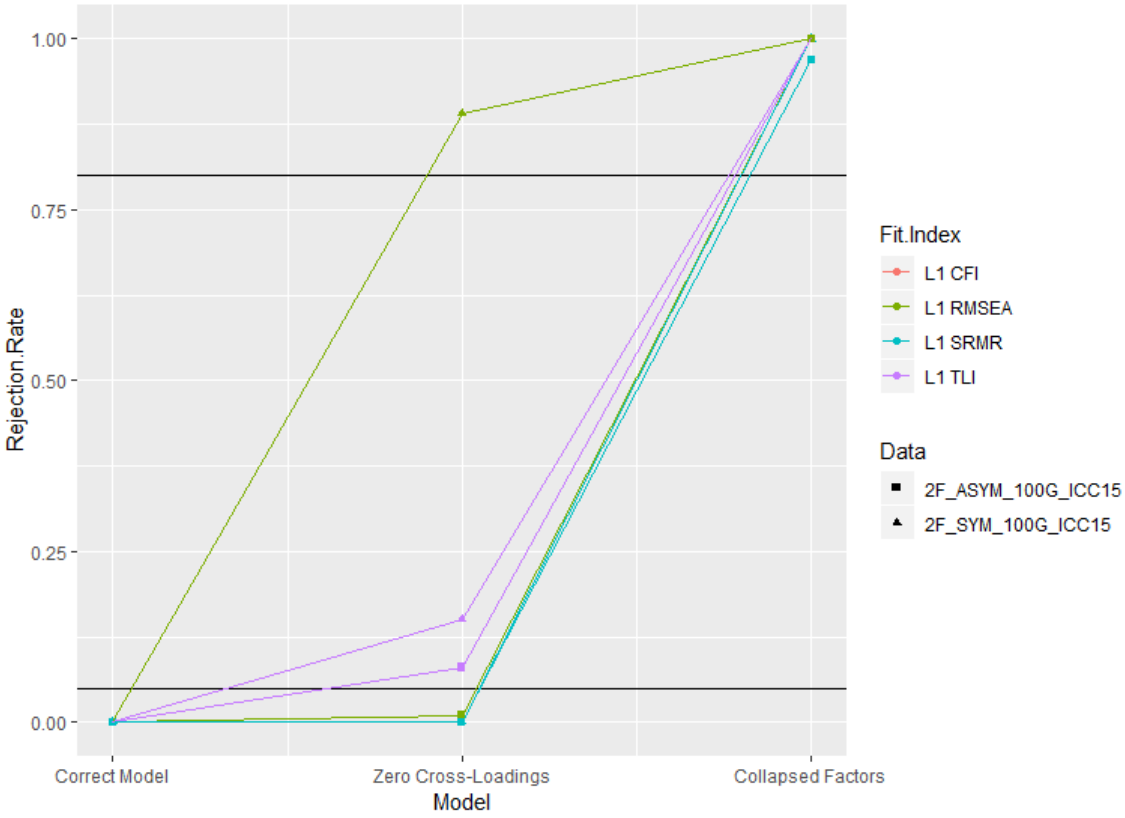

Figure 17. Impact of Data Asymmetry on Level-1 Fit Indices' Rejection Rates with Small Models, Many Groups, and Small ICCs.
Note. This figure has considerable overlap. For correct models, all Level-1 fit indices' rejection rates were zero regardless of asymmetry. With cross-loadings fixed to zero, Level-1 CFI and SRMR rejection rates were zero regardless of asymmetry. For collapsed factors, rejection rates were 1 for Level-1 CFI, TLI, and RMSEA regardless of asymmetry and Level-1 SRMR with symmetry.

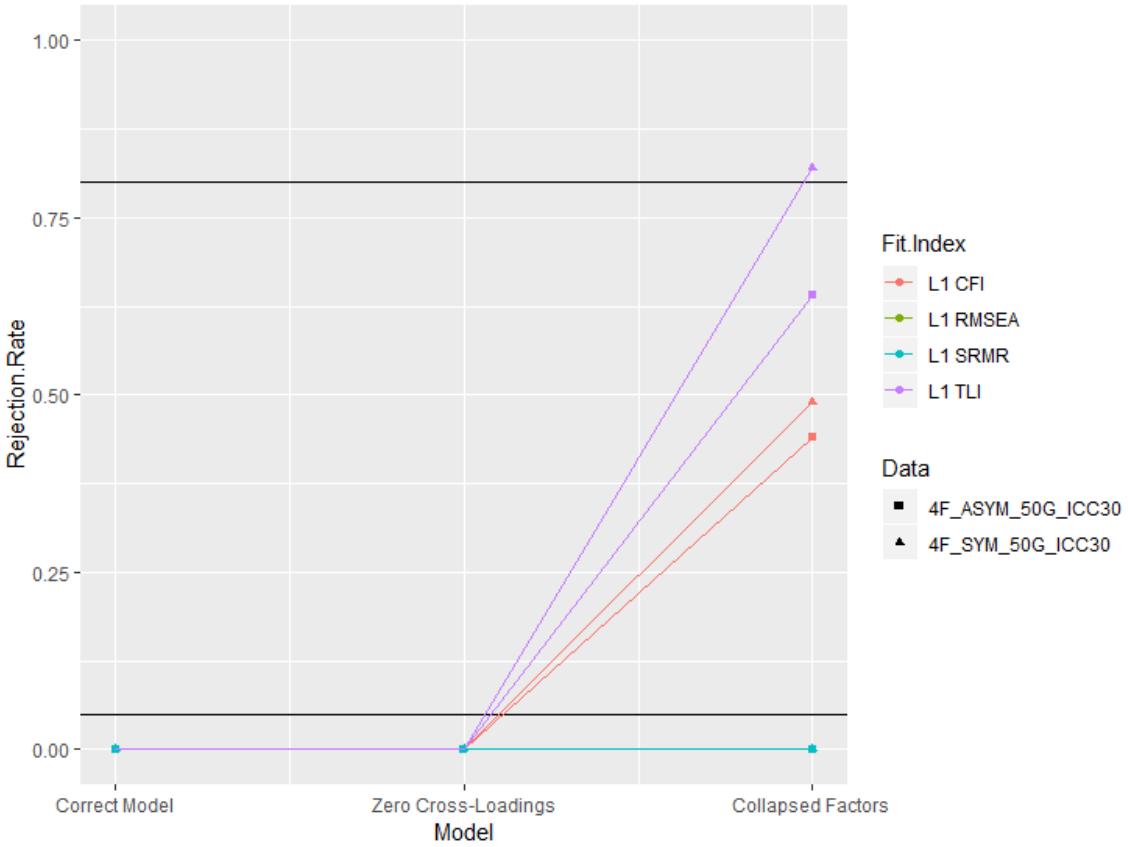

Figure 18. Impact of Data Asymmetry on Level-1 Fit Indices' Rejection Rates with Large Models, Few Groups, and Large ICCs.

Note. This figure has considerable overlap. All Level-1 fit indices' rejection rates were zero for correct models and cross-loadings fixed to zero regardless of asymmetry. For collapsed factors, Level-1 RMSEA and SRMR's rejection rates were zero regardless of asymmetry.

Level-1 fit indices' rejection of Level-1 cross-loadings fixed to 0. Data asymmetry only affected some Level-1 indices with small models and never affected any Level-1 indices with large models. Level-1 CFI and Level-1 SRMR never rejected Level-1 cross-loadings fixed to 0 regardless of data condition. Thus, Level-1 CFI and Level-1 SRMR rejection of Level-1 cross-loadings fixed to 0 was unaffected by data asymmetry. Data asymmetry only affected Level-1 TLI and Level-1 RMSEA's rejection of Level-1 cross-loadings fixed to 0 for small models. Level-1 TLI and Level-1 RMSEA never rejected Level-1 cross-loadings fixed to 0 for large models regardless of asymmetry.

For small models, data asymmetry inconsistently affected Level-1 TLI rejection of Level-1 cross-loadings fixed to 0. Data asymmetry did not affect Level-1 TLI rejection rates with large ICCs. With small ICCs and few groups, Level-1 TLI rejection rates were .11 with symmetric data and .15 with severely asymmetric data. With small ICCs and many groups, Level-1 TLI rejection rates were .15 with symmetric data and .08 with severely asymmetric data.

For small models, Level-1 RMSEA rejection of Level-1 cross-loadings fixed to 0 always decreased as asymmetry increased. Rejection rates were .09 to .89 for symmetric data and 0 to .05 for severely asymmetric data. The largest decreases occurred with small ICCs. Level-1 RMSEA rejected 73% to 89% of small models fit to symmetric data with small ICCs and 1% to 5% of small models fit to severely asymmetric data with small ICCs.

Level-1 fit indices' rejection of Level-1 collapsed factors. Generally, Level-1 CFI rejection of Level-1 collapsed factors decreased as asymmetry increased. Rejection rates were .49 to 1 for symmetric data and .44 to 1 for severely asymmetric data. For symmetric data, Level-1 CFI performed worst with large models, few groups, and large ICCs (rejection rate = .49). Otherwise, with symmetric data, Level-1 CFI rejected 82% to 100% of Level-1 collapsed factors. For asymmetric data, Level-1 CFI performed best with small models (rejection rates = .82 to 1). Level-1 CFI rejected 100% of Level-1 collapsed factors regardless of asymmetry with small models, many groups, and small ICCs.

Level-1 TLI rejection of Level-1 collapsed factors decreased as asymmetry increased or remained at 100% regardless of asymmetry. For symmetric data, Level-1 TLI rejected 82% to 100% of Level-1 collapsed factors. With symmetric data, Level-1 TLI performed worst with large models, few groups, and large ICCs (rejection rate = .82). Otherwise, Level-1 TLI rejected 99% to 100% of Level-1 collapsed factors fit to symmetric data. Data asymmetry usually did not affect Level-1 TLI with small models. Level-1 TLI rejected 100% of small collapsed models with symmetric data and 98% to 100% with severely asymmetric data. With large models, Level-1 TLI rejection rates mostly decreased as asymmetry increased. Rejection rates were .82 to 1 with data symmetry and .64 to .97 with asymmetry. With asymmetric data, Level-1 TLI performed worst with large models, few groups, and large ICCs (rejection rate = .64). Otherwise, with asymmetric data, Level-1 TLI rejected at least 88% of collapsed models.

Data asymmetry inconsistently affected Level-1 RMSEA rejection of Level-1 collapsed factors. Level-1 RMSEA's rejection of Level-1 collapsed factors either decreased as asymmetry increased or did not change. Rejection rates were 0 to 1 for symmetric data and 0 to 1 for severely asymmetric data. Level-1 RMSEA never rejected large collapsed models with large ICCs regardless of data asymmetry (rejection rate = 0). Level-1 RMSEA always rejected small collapsed models with symmetric data (rejection rate = 1).

Data asymmetry inconsistently affected Level-1 SRMR rejection of Level-1 collapsed factors. For large collapsed models, Level-1 SRMR rejection rates generally did not change as asymmetry increased. For large collapsed models, Level-1 SRMR rejection rates were 0 with symmetric data and 0 to .04 with severely asymmetric data. For small collapsed models, Level-1 SRMR rejection rates generally increased as asymmetry increased. Rejection rates were .54 to 1 with symmetric data and .67 to .94 with severely asymmetric data. With small models, many groups, and large ICCs, Level-1 SRMR yielded rejection rates of .54 with symmetric data and .67 with asymmetric data. Level-1 SRMR rejection of Level-1 collapsed factors decreased slightly with small models, many groups, and small ICCs. In this situation, Level-1 SRMR rejected 100% of Level-1 collapsed factors with symmetric data and 97% with severely asymmetric data.

Level-1 fit indices' rejection of correct Level-1 model. Level-1 CFI, TLI, RMSEA, and SRMR never rejected the correct Level-1 model.

Model Size's Impact on Level-1 Fit Indices

Before describing the impact of model size on Level-1 fit indices' rejection rates, figures will be shown. These figures introduce the complex nature of level-specific fit index performance. Model size's impact generally also depended on the specific fit index, type of misspecification, data asymmetry, number of groups, and ICCs. Figures 19 through 22 contrast Level-1 fit indices' rejection rates with small and large models for certain conditions. Tables 4 and 5 give Level-1 fit indices' rejection rates.

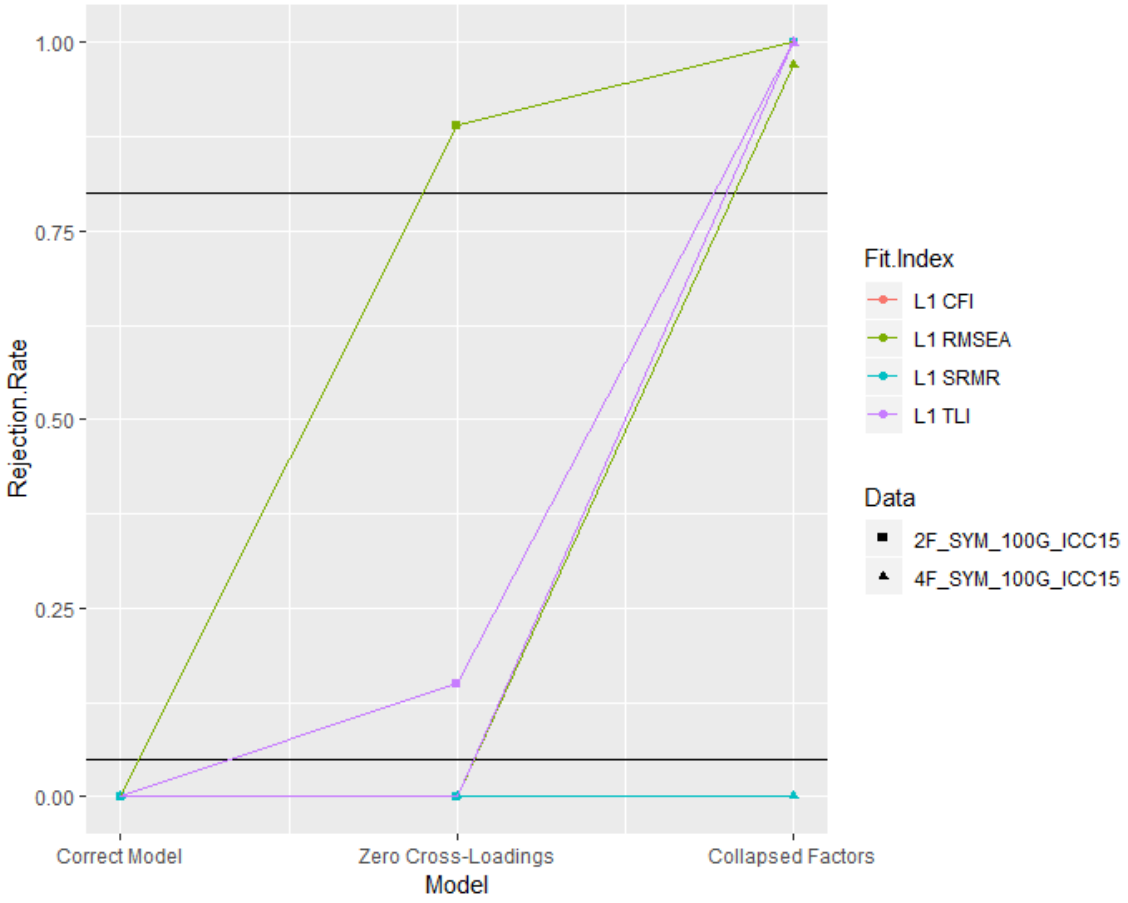

Figure 19. Impact of Model Size on Level-1 Fit Indices' Rejection of Symmetric Data, Many Groups, and Small ICCs.

Note. This figure has considerable overlap. All Level-1 fit indices' rejection rates were zero for correct models regardless of model size. For zero cross-loadings, rejection rates were zero for Level-1 CFI and SRMR with small models and all Level-1 fit indices with large models. With collapsed factors, rejection rates were 1 for all Level-1 fit indices with small models and Level-1 CFI and TLI with large models.

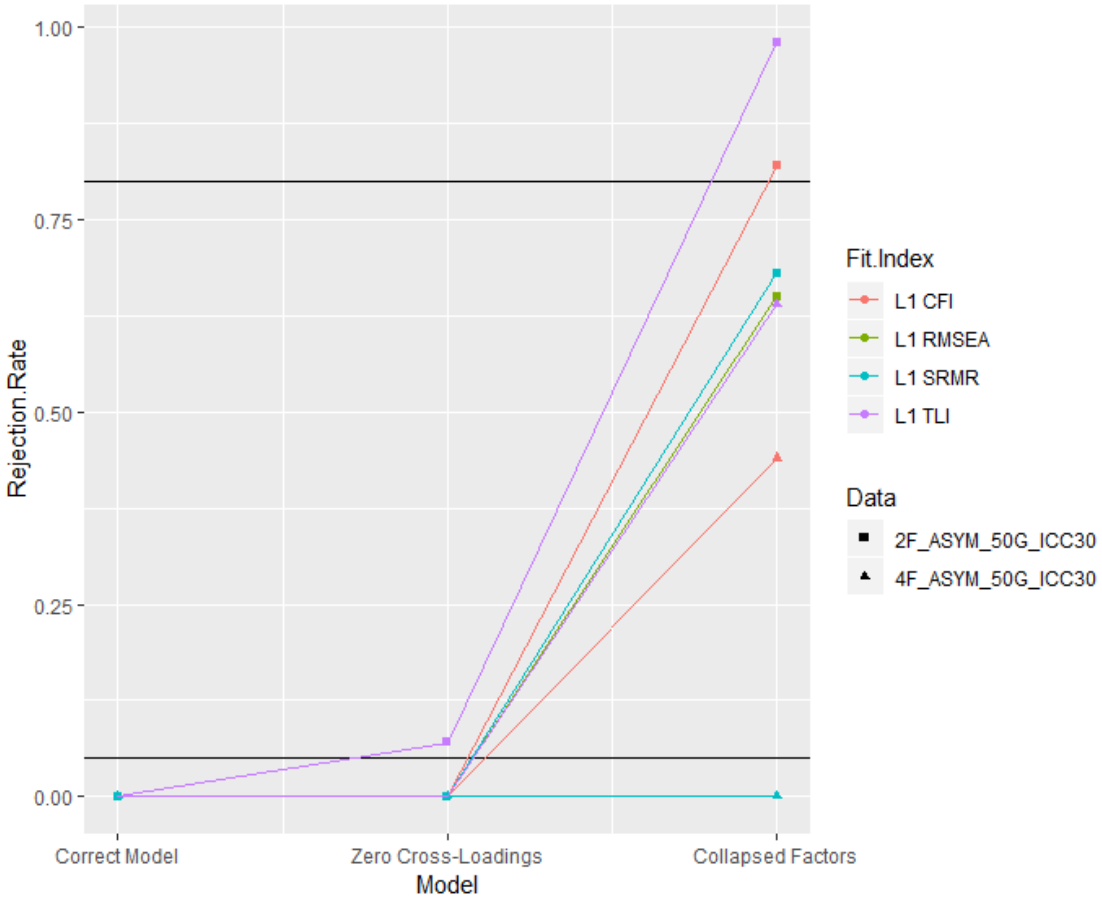

Figure 20. Impact of Model Size on Level-1 Fit Indices' Rejection of Asymmetric Data, Few Groups, and Large ICCs.

Note. This figure has considerable overlap. For correct models, all Level-1 fit indices' rejection rates were zero regardless of model size. For zero cross-loadings, rejection rates were zero for Level-1 CFI, RMSEA, and SRMR regardless of model size and Level-1 TLI with large models. With collapsed factors, Level-1 RMSEA and SRMR's rejection rates were zero for large models.

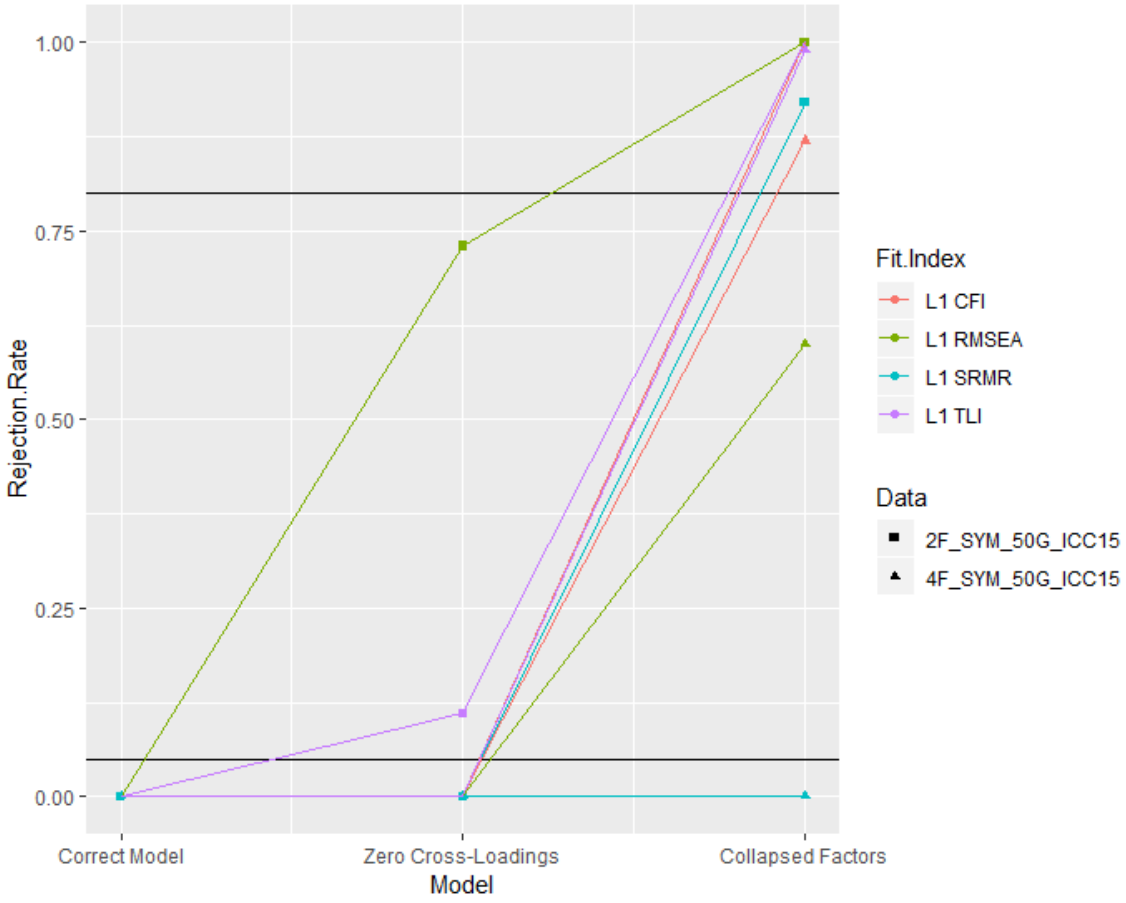

Figure 21. Impact of Model Size on Level-1 Fit Indices' Rejection of Symmetric Data, Few Groups, and Small ICCs.

Note. This figure has considerable overlap. For correct models, all Level-1 fit indices' rejection rates were zero regardless of model size. For zero cross-loadings, rejection rates were zero for all Level-1 fit indices with large models and Level-1 CFI and SRMR with small models. With small models, Level-1 CFI, TLI, and RMSEA's rejection rates were 1 for collapsed factors.

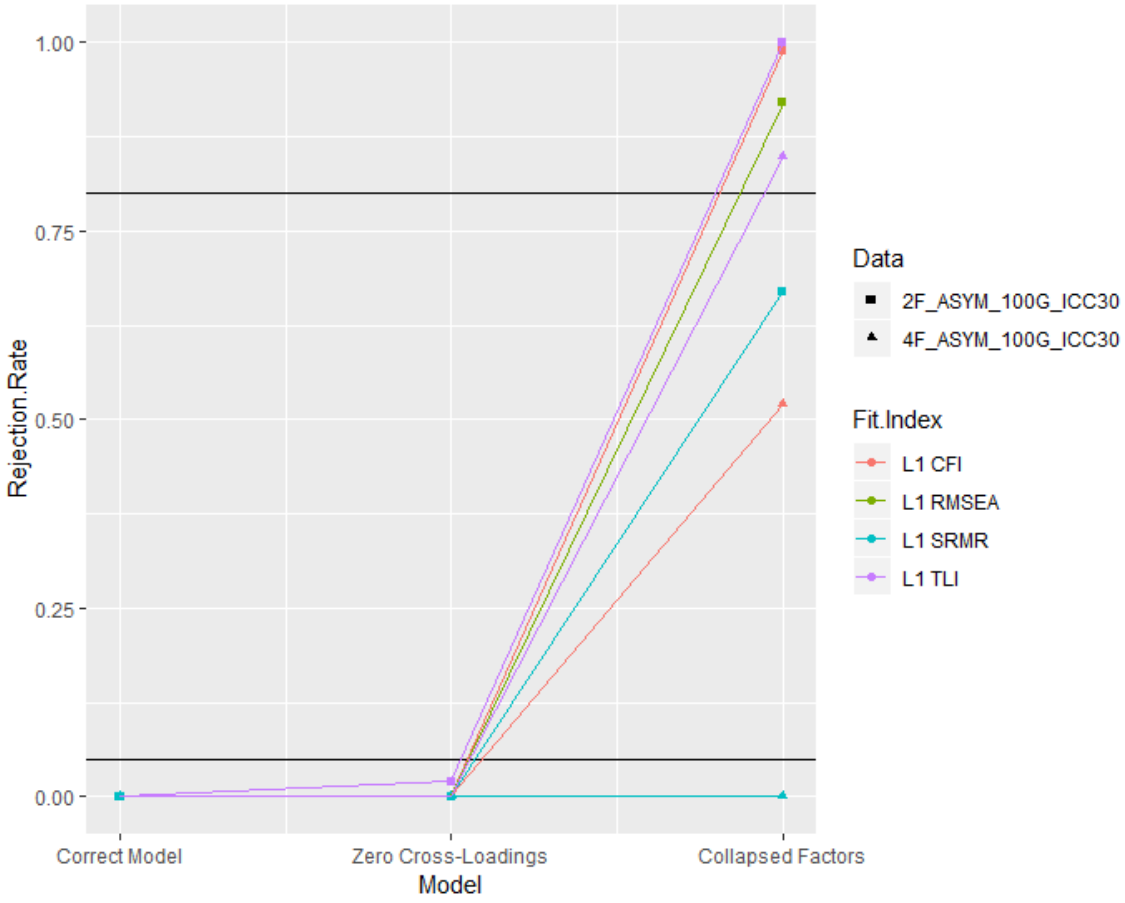

Figure 22. Impact of Model Size on Level-1 Fit Indices' Rejection of Asymmetric Data, Many Groups, and Large ICCs.

Note. This figure has considerable overlap. For correct models, all Level-1 fit indices' rejection rates were zero regardless of model size. For zero cross-loadings, rejection rates were zero for Level-1 CFI, RMSEA, and SRMR with small models and all Level-1 fit indices with small models. For collapsed factors, rejection rates were zero for Level-1 RMSEA and SRMR with large models.

Level-1 fit indices' rejection of Level-1 cross-loadings fixed to 0. Model size only impacted Level-1 TLI and RMSEA's rejection of Level-1 cross-loadings fixed to 0. Level-1 CFI and Level-1 SRMR never rejected Level-1 cross-loadings fixed to 0 regardless of model size. Level-1 TLI rejection rates for Level-1 cross-loadings fixed to 0 always decreased as model size increased. Rejection rates were .02 to .15 for small models and 0 for large models. Level-1 RMSEA rejection of Level-1 cross-loadings fixed to 0 usually decreased as model size increased. Rejection rates were 0 to .89 for small models and 0 for large models. For small models, Level-1 RMSEA performed worst with severe asymmetry (rejection rates = 0 to .05). Level-1 RMSEA also performed poorly with symmetric data and large ICCs (rejection rates = .09 to .16). Level-1 RMSEA performed well only with small models fit to symmetric data with small ICCs (rejection rates = .73 to .89).

Level-1 fit Indices' rejection of collapsed Level-1 factors. Level-1 CFI rejection of Level-1 collapsed factors usually decreased as model size increased. Rejection rates were .82 to 1 for small models and .44 to 1 for large models. With small models, Level-1 CFI performed worst with severe asymmetry, few groups, and large ICCs (rejection rate = .82). Otherwise, Level-1 CFI rejected 94% to 100% of small collapsed models. With large models, Level-1 CFI performed worst with asymmetric data, few groups, and large ICCs (rejection rate = .44). With large models, Level-1 CFI performed best with symmetric data, many groups, and small ICCs (rejection rate = 1.0).

Level-1 TLI rejection of Level-1 collapsed factors usually decreased as model size increased. Rejection rates were .98 to 1 for small models and .64 to 1 for large

models. The largest decrease (.34) occurred with severe asymmetry, few groups, and large ICCs. In this situation, Level-1 TLI rejected 98% of small models and 64% of large models. Otherwise, Level-1 TLI performed well with large models, yielding rejection rates of .82 to 1. Level-1 TLI performed better with large collapsed models than Level-1 CFI performed.

Level-1 RMSEA rejection of Level-1 collapsed factors always decreased as model size increased. Rejection rates were .65 to 1 for small models and 0 to .97 for large models. With small models, Level-1 RMSEA performed worst with severe asymmetry, few groups, and small ICCs (rejection rate = .65). Otherwise, Level-1 RMSEA rejected 92% to 100% of small collapsed models. With large models, Level-1 RMSEA usually never rejected Level-1 collapsed factors. However, with symmetry and small ICCs, Level-1 RMSEA rejected 60% of large models with few groups and 97% with many groups.

Level-1 SRMR rejection of Level-1 collapsed factors always decreased as model size increased. Rejection rates were .54 to 1 for small models and 0 to .04 for large models. For small models, Level-1 SRMR performed worst with large ICCs, yielding rejection rates of .54 to .67. Otherwise, Level-1 SRMR rejected 92% to 100% of small collapsed models.

Level-1 fit indices' rejection of correct Level-1 model. Level-1 CFI, TLI, RMSEA, and SRMR never rejected the correct Level-1 model regardless of model size.

Number of Groups' Impact on Level-1 Fit Indices

Before describing the impact of number of groups on Level-1 fit indices' rejection, figures will be shown. These figures introduce the complex nature of level-specific fit index performance. Number of groups' impact generally also depended on the specific fit index, type of misspecification, data asymmetry, model size, and ICCs. Figures 23 through 25 contrast Level-1 fit indices' rejection rates with few and many groups for various combinations of conditions. Recall that Tables 4 and 5 provide rejection rates for Level-1 fit indices.

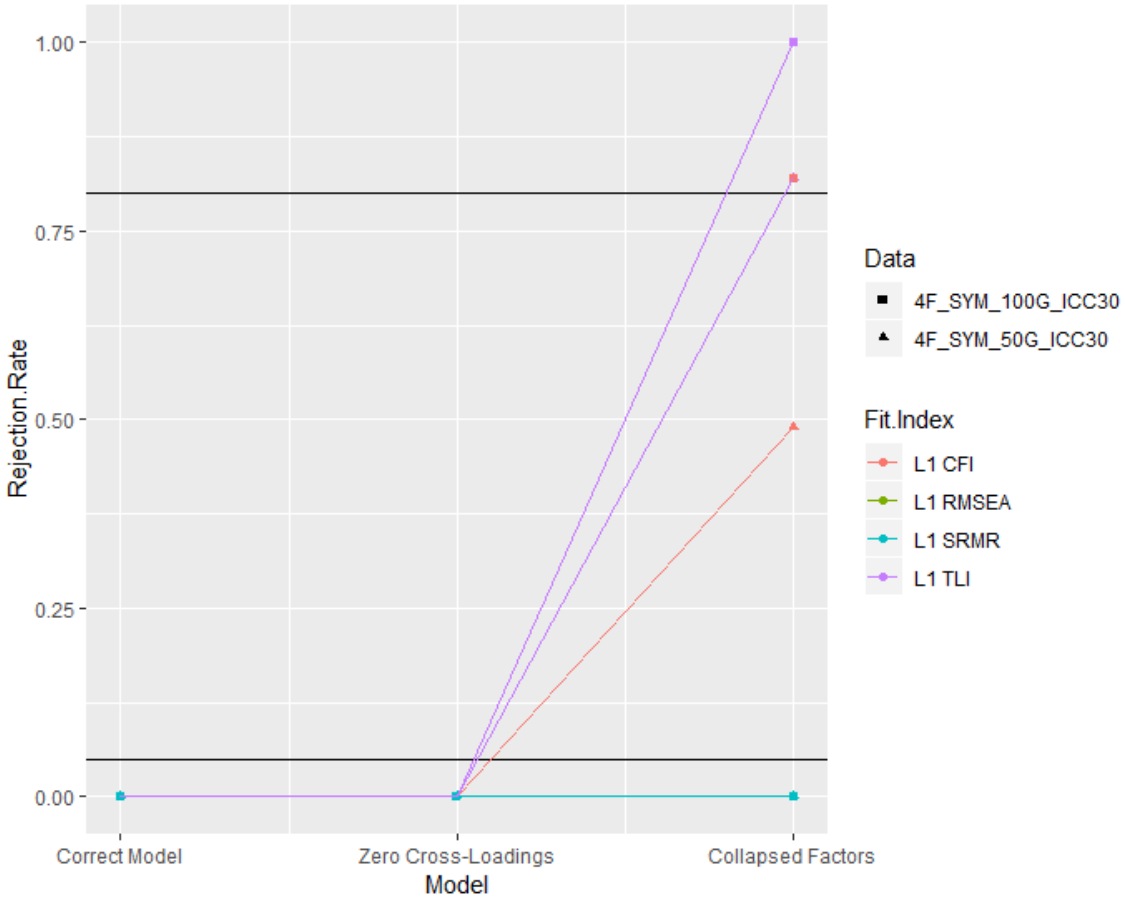

Figure 23. Impact of Number of Groups on Level-1 Fit Indices' Rejection of Large Models, Symmetric Data, and Large ICCs.

Note. This figure has considerable overlap. All Level-1 fit indices had rejection rates of 0 for correct models and zero cross-loadings regardless of number of groups. For collapsed factors, Level-1 RMSEA and SRMR had rejection rates of zero regardless of number of groups. For collapsed factors, Level-1 CFI's rejection rate with many groups was identical to Level-1 TLI with few groups.

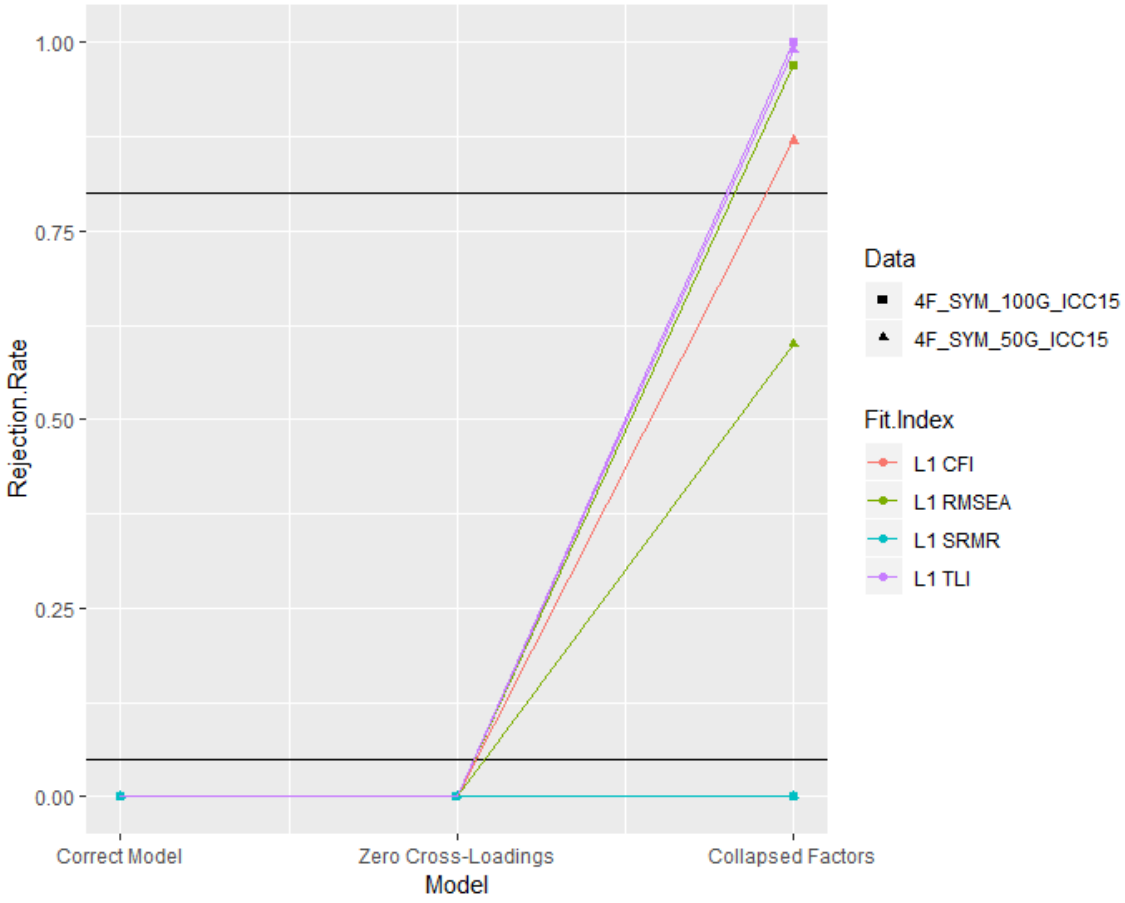

Figure 24. Impact of Number of Groups on Level-1 Fit Indices' Rejection of Large Models, Symmetric Data, and Small ICCs.

Note. This figure has considerable overlap. For correct models and zero cross-loadings, all Level-1 fit indices' rejection rates were zero. For collapsed factors and many groups, Level-1 CFI and TLI had identical rejection rates of 1. For collapsed factors, Level-1 SRMR had rejection rates of zero regardless of number of groups.

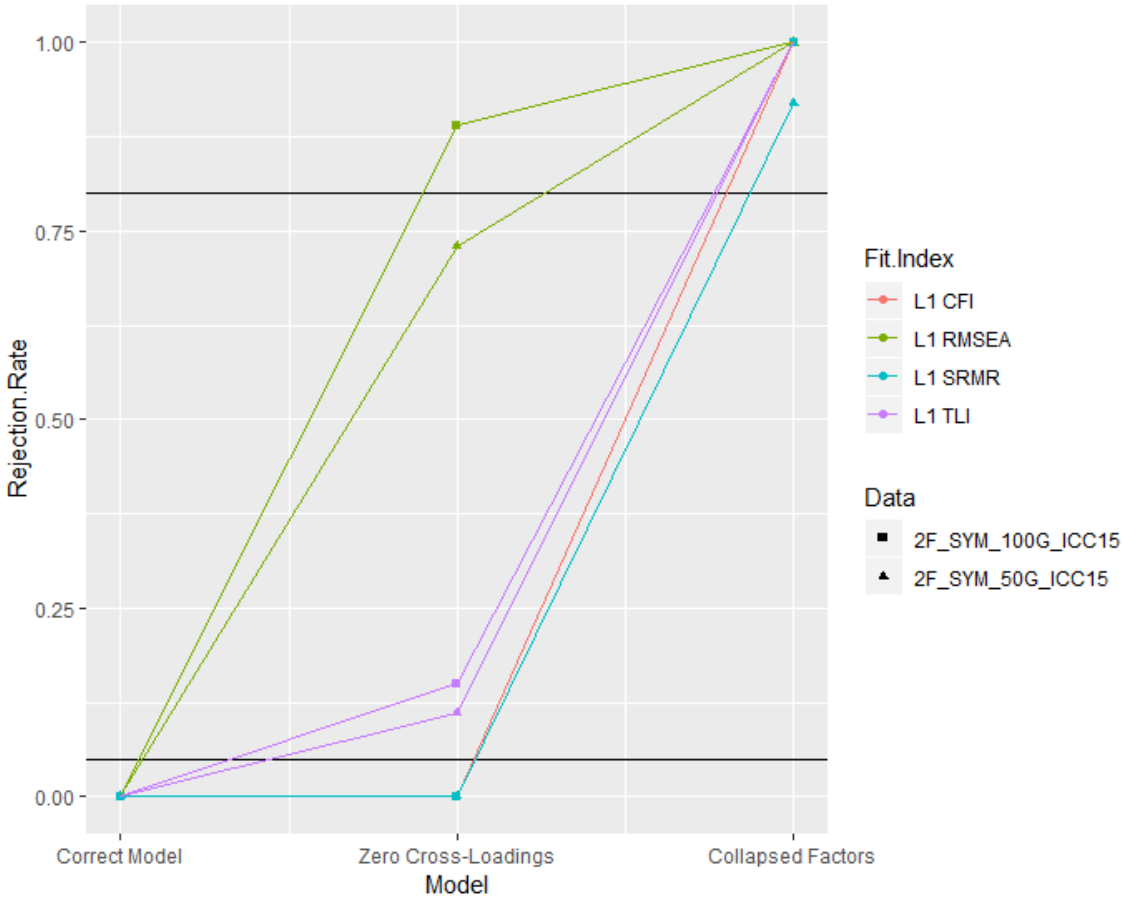

Figure 25. Impact of Number of Groups on Level-1 Fit Indices' Rejection of Small Models, Symmetric Data, and Small ICCs.

Note. This figure has considerable overlap. For correct models, all Level-1 fit indices had rejection rates of zero regardless of number of groups. For zero cross-loadings, Level-1 CFI and SRMR rejection rates were zero regardless of number of groups. For collapsed factors, rejection rates were 1 for Level-1 CFI, TLI, and RMSEA regardless of number of groups and Level-1 SRMR 1 with many groups.

Level-1 fit indices' rejection of Level-1 cross-loadings fixed to 0. Number of groups only affected Level-1 TLI and RMSEA rejection of Level-1 cross-loadings fixed to 0 for small models, but it never affected Level-1 CFI and SRMR. Level-1 CFI and SRMR never rejected Level-1 cross-loadings fixed to 0 regardless of number of groups. With small models, Level-1 TLI rejection of Level-1 cross-loadings fixed to 0 usually decreased as number of groups increased. Rejection rates were .07 to .15 with few groups and .02 to .15 with many groups. For small models, Level-1 TLI rejection rates did increase as number of groups increased with symmetry and small ICCs. In this situation, Level-1 TLI rejection rates were .11 with few groups and .15 with many groups.

Number of groups inconsistently affected Level-1 RMSEA rejection of cross-loadings fixed to 0. With small models and symmetric data, Level-1 RMSEA rejection rates increased as number of groups increased. Rejection rates were .09 to .73 with few groups and .16 to .89 with many groups. Level-1 RMSEA performed best with small models, symmetric data, and small ICCs. In this situation, Level-1 RMSEA rejected 73% of Level-1 cross-loadings fixed to 0 with few groups and 89% with many groups. Level-1 RMSEA rejection rates of other small models were always zero regardless of number of groups (asymmetry and large ICCs) or decreased with asymmetry and small ICCs.

Level-1 fit indices' rejection of Level-1 collapsed factors. Level-1 CFI rejection of Level-1 collapsed factors usually increased as number of groups increased. Rejection rates were .44 to 1 with few groups and .52 to 1 with many groups. Level-1 CFI rejected 100% of Level-1 collapsed factors regardless of number of groups when fitting a small model to symmetric data with small ICCs. Level-1 CFI performed best with small

models, producing rejection rates of .82 to 1 with few groups and .99 to 1 with many groups. Level-1 CFI performed worst with large models, severe asymmetry, and large ICCs, yielding rejection rates of .44 with few groups and .49 with many groups.

Level-1 TLI rejection of Level-1 collapsed factors usually increased as number of groups increased. Rejection rates were .64 to 1 with few groups and .85 to 1 with many groups. Level-1 TLI performed worst with large models fit to severely asymmetric data with large ICCs. In this situation, Level-1 TLI rejected 64% of collapsed models with few groups and 87% with many groups. Otherwise, Level-1 TLI rejection rates were .82 to 1 with few groups and .97 to 1 with many groups. Level-1 TLI rejected 100% of small models regardless of number of groups with symmetric data or asymmetric data with small ICCs.

Level-1 RMSEA rejection of Level-1 collapsed factors usually did not depend on number of groups. With small models, Level-1 RMSEA usually rejected 100% of Level-1 collapsed factors regardless of number of groups. However, when fitting small models to severely asymmetric data with large ICCs, Level-1 RMSEA rejection rates increased from .65 to .92 as number of groups increased. With large models, Level-1 RMSEA usually never rejected Level-1 collapsed factors regardless of number of groups. When fitting large models to symmetric data with small ICCs, Level-1 RMSEA rejection rates increased from .60 to .97 as number of groups increased.

Number of groups usually did not affect Level-1 SRMR rejection of large collapsed models and inconsistently affected Level-1 SRMR's rejection of small collapsed models. With large models, Level-1 SRMR usually never rejected Level-1

collapsed factors regardless of number of groups. When fitting large models to severely asymmetric data with small ICCs, Level-1 SRMR rejection rates decreased from .04 with few groups to 0 with many groups. For small models and small ICCs, Level-1 SRMR rejection rates increased as number of groups increased. Rejection rates were .92 to .94 with few groups and .97 to 1 with many groups. For small models and large ICCs, Level-1 SRMR rejection rates decreased as number of groups increased. These situations produced rejection rates of .62 to .68 with few groups and .54 to .67 with many groups.

Level-1 fit indices' rejection of correct Level-1 model. Level-1 CFI, TLI, RMSEA, and SRMR never rejected the correct Level-1 model regardless of number of groups.

ICCs' Impact on Level-1 Fit Indices

Before describing ICCs' impact on Level-1 fit indices' rejection rates, figures will be shown. These figures introduce the complex nature of level-specific fit index performance. ICCs' impact generally also depended on the specific fit index, type of misspecification, model size, number of groups, and data asymmetry. Figures 26 and 27 contrast Level-1 fit indices' rejection rates with small and large ICCs for various combinations of conditions. Recall that Tables 4 and 5 provide all rejection rates for Level-1 fit indices.

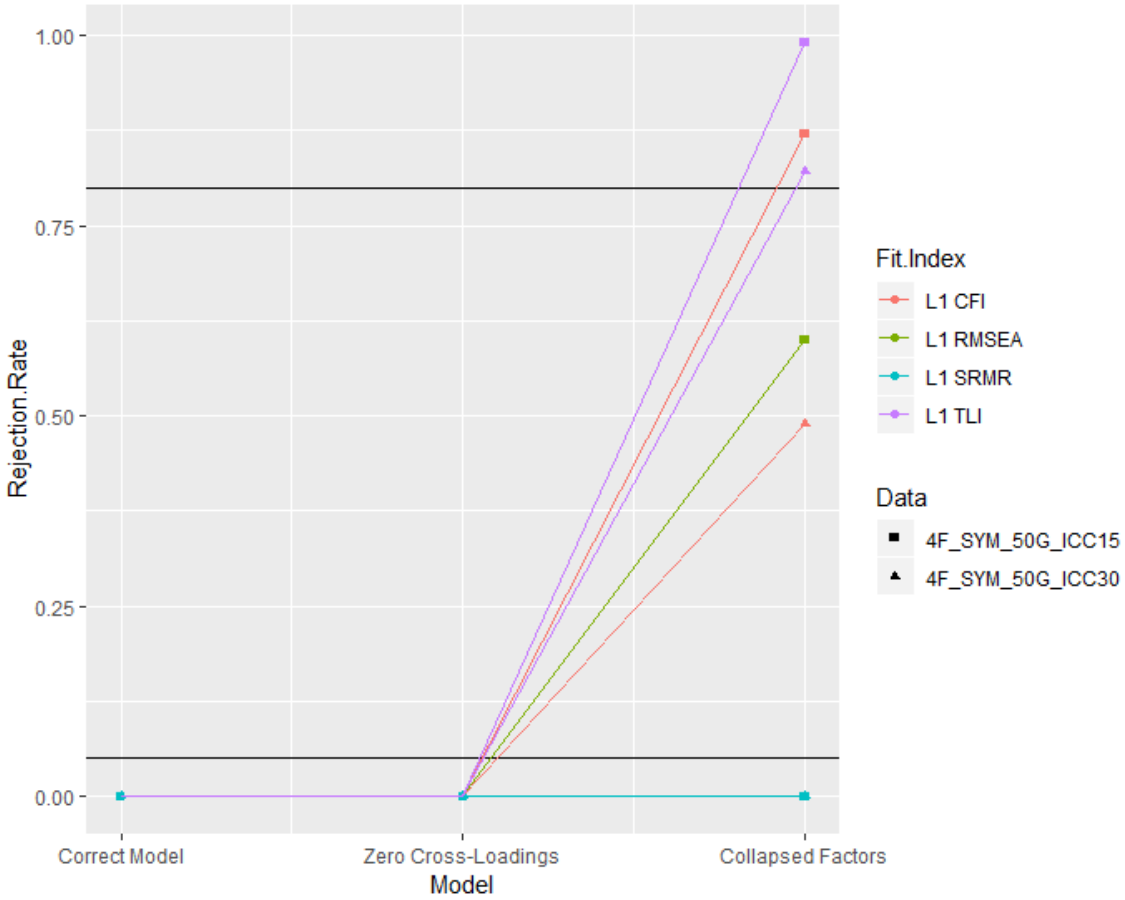

Figure 26. Impact of ICCs on Level-1 Fit Indices' Rejection Rates with Large Models, Symmetric Data, and Few Groups.

Note. This figure has considerable overlap. All Level-1 fit indices' rejection rates were zero for correct models and cross-loadings fixed to zero regardless of ICCs. For collapsed factors, rejection rates were zero for Level-1 RMSEA and SRMR with large ICCs and for Level-1 SRMR with small ICCs.

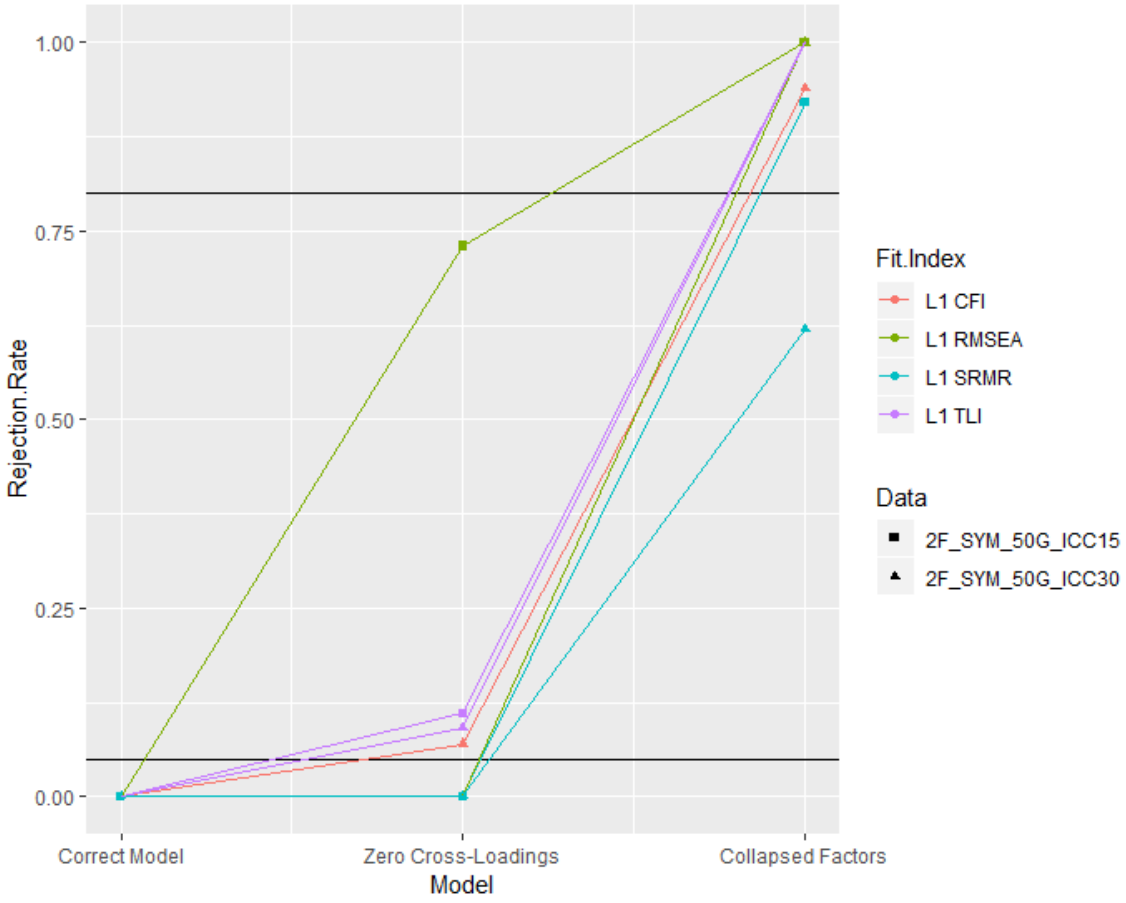

Figure 27. Impact of ICCs on Level-1 Fit Indices' Rejection Rates with Small Models, Symmetric Data, and Few Groups.

Note. This figure has considerable overlap. All Level-1 fit indices' rejection rates were zero for correct models regardless of ICCs. Level-1 CFI and SRMR's rejection rates were zero for cross-loadings fixed to zero regardless of ICCs. For collapsed factors, rejection rates were 1 for Level-1 CFI, TLI, and RMSEA with small ICCs and Level-1 TLI and RMSEA with large ICCs.

Level-1 fit indices' rejection of Level-1 cross-loadings fixed to 0. ICCs never affected Level-1 CFI and SRMR rejection of Level-1 cross-loadings fixed to 0, and affected Level-1 TLI and RMSEA only with small models. Level-1 CFI and SRMR never rejected Level-1 cross-loadings fixed to 0 regardless of ICCs. When fitting small models, Level-1 TLI and RMSEA rejection rates always decreased as ICCs increased. Level-1 TLI rejection rates were .08 to .15 with small ICCs and .02 to .07 with large ICCs. Level-1 RMSEA rejection rates were .01 to .89 with small ICCs and 0 to .16 with large ICCs. Level-1 RMSEA performed best with small models, symmetric data, and small ICCs, yielding rejection rates of .73 to .89.

Level-1 fit indices' rejection of Level-1 collapsed factors. Level-1 CFI rejection of Level-1 collapsed factors usually decreased as ICCs increased. Rejection rates were .64 to 1 with small ICCs and .44 to 1 with large ICCs. Level-1 CFI performed best with small models, producing rejection rates of .98 to 1 with small ICCs and .82 to 1 with large ICCs. Level-1 CFI performance deteriorated greatly as ICCs increased for all large model conditions. With large models, Level-1 CFI rejection rates were .64 to 1 with small ICCs and .44 to .82 with large ICCs. With large models, Level-1 CFI performed best with symmetric data and many groups, yielding rejection rates of 1 with small ICCs and .82 with large ICCs. Otherwise, Level-1 CFI rejection rates were .44 to .52 with large models and large ICCs.

Level-1 TLI rejection of Level-1 collapsed factors decreased as ICCs increased or remained 100% regardless of ICCs. Rejection rates were .88 to 1 with small ICCs and .64 to 1 with large ICCs. Level-1 TLI performed best with small models, producing rejection

rates of 1 with small ICCs and .98 to 1 with large ICCs. Level-1 TLI performed worst with large models, severe asymmetry, and few groups. In this situation, Level-1 TLI rejection rates were .88 with small ICCs and .64 with large ICCs. Otherwise, Level-1 TLI rejection rates were .97 to 1 with small ICCs and .82 to 1 with large ICCs.

ICCs usually did not affect Level-1 RMSEA rejection of Level-1 collapsed factors. Level-1 RMSEA never rejected large models fit to severely asymmetric data regardless of ICCs. Level-1 RMSEA always rejected small models fit to symmetric data regardless of ICCs. With small ICCs, Level-1 RMSEA always rejected small models. Level-1 RMSEA rejection of Level-1 collapsed factors decreased as ICCs increased for large models fit to symmetric data or small models fit to asymmetric data. Rejection rates were .6 to 1 with small ICCs and 0 to .92 with large ICCs. The largest decreases in Level-1 RMSEA performance due to increasing ICCs occurred with large models fit to symmetric data. In these situations, Level-1 RMSEA rejected 60% (few groups) to 97% (many groups) with small ICCs, but 0% with large ICCs.

Level-1 SRMR rejection of Level-1 collapsed factors usually decreased as ICCs increased. Rejection rates were .04 to 1 with small ICCs and 0 to .54 with large ICCs. Level-1 SRMR usually never rejected large collapsed models regardless of ICCs. However, with severe asymmetry and few groups, Level-1 SRMR rejection rates were .04 with small ICCs and 0 with large ICCs. With small models, Level-1 SRMR rejection rates were .92 to 1 with small ICCs and .54 to .68 with large ICCs.

Level-1 fit indices' rejection of correct Level-1 model. Level-1 CFI, TLI, RMSEA, and SRMR never rejected the correct Level-1 model regardless of ICCs.

Level-2 Fit Indices' Overall Performance

Tables 6 and 7 provide rejection rates for Level-2 fit indices.

Table 6. Rejection Rates of Level-2 Fit Indices for Small Model Conditions

Fit Index	L2 Misp. Cross-loadings	L2 Misp. Factors	Correct Model	Data
L2 CFI	0.45	0.82	0.22	Skewed data, 50 groups, ICC = .15
L2 TLI	0.52	0.82	0.26	
L2 RMSEA	0.39	0.78	0.2	
L2 SRMR	0.32	0.68	0.09	
L2 CFI	0.57	0.97	0.15	Skewed data, 100 groups, ICC = .15
L2 TLI	0.65	0.97	0.27	
L2 RMSEA	0.46	0.9	0.1	
L2 SRMR	0.01	0.41	0	
L2 CFI	0.5	0.91	0.16	Skewed data, 50 groups, ICC = .30
L2 TLI	0.59	0.92	0.18	
L2 RMSEA	0.49	0.88	0.15	
L2 SRMR	0.11	0.6	0	
L2 CFI	0.67	1	0.04	Skewed data, 100 groups, ICC = .30
L2 TLI	0.77	1	0.17	
L2 RMSEA	0.69	1	0.08	
L2 SRMR	0	0.41	0	
L2 CFI	0.52	0.87	0.18	Normal data, 50 groups, ICC = .15
L2 TLI	0.56	0.88	0.26	
L2 RMSEA	0.49	0.84	0.16	
L2 SRMR	0.17	0.55	0.01	
L2 CFI	0.67	1	0.13	Normal data, 100 groups, ICC = .15
L2 TLI	0.76	1	0.28	
L2 RMSEA	0.6	0.98	0.12	
L2 SRMR	0.02	0.39	0	
L2 CFI	0.46	0.91	0.07	Normal data, 50 groups, ICC = .30
L2 TLI	0.55	0.94	0.13	
L2 RMSEA	0.48	0.91	0.07	
L2 SRMR	0.05	0.51	0	
L2 CFI	0.81	1	0.03	Normal data, 100 groups, ICC = .30
L2 TLI	0.87	1	0.12	
L2 RMSEA	0.84	1	0.06	
L2 SRMR	0	0.42	0	

Note. Misp = misspecified.

Table 7. Rejection Rates of Level-2 Fit Indices for Large Model Conditions

Fit Index	L2 Misp. Cross-loadings	L2 Collapsed Factors	Correct Model	Data
L2 CFI	0.33	0.75	0.16	Skewed data, 50 groups, ICC = .15
L2 TLI	0.35	0.77	0.2	
L2 RMSEA	0.05	0.24	0.01	
L2 SRMR	0.89	0.99	0.8	
L2 CFI	0.4	0.9	0.15	Skewed data, 100 groups, ICC = .15
L2 TLI	0.43	0.93	0.15	
L2 RMSEA	0.02	0.45	0	
L2 SRMR	0.09	0.79	0.01	
L2 CFI	0.2	0.84	0.07	Skewed data, 50 groups, ICC = .30
L2 TLI	0.24	0.85	0.07	
L2 RMSEA	0.05	0.5	0	
L2 SRMR	0.41	0.95	0.08	
L2 CFI	0.43	1	0.03	Skewed data, 100 groups, ICC = .30
L2 TLI	0.54	1	0.04	
L2 RMSEA	0.06	0.89	0	
L2 SRMR	0	0.53	0	
L2 CFI	0.33	0.81	0.2	Normal data, 50 groups, ICC = .15
L2 TLI	0.4	0.85	0.22	
L2 RMSEA	0.04	0.41	0.01	
L2 SRMR	0.67	0.96	0.34	
L2 CFI	0.49	0.98	0.08	Normal data, 100 groups, ICC = .15
L2 TLI	0.57	0.99	0.13	
L2 RMSEA	0.03	0.62	0	
L2 SRMR	0	0.54	0	
L2 CFI	0.16	0.83	0.01	Normal data, 50 groups, ICC = .30
L2 TLI	0.2	0.83	0.01	
L2 RMSEA	0.03	0.48	0	
L2 SRMR	0.24	0.87	0	
L2 CFI	0.47	1	0.02	Normal data, 100 groups, ICC = .30
L2 TLI	0.54	1	0.04	
L2 RMSEA	0.11	0.93	0	
L2 SRMR	0	0.36	0	

Note. Misp = misspecified.

Level-2 fit indices' rejection of Level-2 cross-loadings fixed to 0. Generally, Level-2 fit indices' power to detect Level-2 cross-loadings fixed to 0 was very low but depended on model size. Level-2 CFI, TLI, and RMSEA had power levels less than .80 except with small models, symmetric data, many groups, and large ICCs. This situation yielded rejection rates of .81 for Level-2 CFI, .87 for Level-2 TLI, and .84 for Level-2 RMSEA. Level-2 CFI rejection rates were .16 to .49 with large models and .45 to .67 for small models. Level-2 TLI's power ranged from 0.20 to 0.54 for large models and .52 to .77 for small models. Both Level-2 CFI and TLI performed worst with large models, symmetric data, few groups, and large ICCs (rejection rate = .16 and .20, respectively). Level-2 RMSEA's power was .04 to .11 for large models and .39 to .69 for small models. For small models, Level-2 RMSEA performed worst with asymmetric data, few groups, and small ICCs (rejection rate = .39). Level-2 SRMR had low power in all conditions except with large models, few groups, and small ICCs. These situations yielded rejection rates of .67 with symmetric data and .89 with asymmetric data. Otherwise, Level-2 SRMR's power was 0 to .32 and 0 for most conditions. As described below, high power SRMR conditions usually were accompanied by extreme Type I error rates.

Level-2 fit indices' rejection of Level-2 collapsed factors. Level-2 fit indices' power to reject Level-2 collapsed factors varied widely across indices. Level-2 CFI and TLI had high power to detect Level-2 collapsed factors in almost all conditions. Level-2 CFI and TLI's power was .75 and .77, respectively, with large model, asymmetric data, few groups, and small ICCs. Otherwise, Level-2 CFI and TLI's power ranged from .82 to

1. Level-2 CFI and TLI performed best with small models, symmetric data, and many groups (rejection rates = 1).

Level-2 RMSEA had much higher power for small models than large models. For small models, Level-2 RMSEA power always exceeded .80 except with asymmetric data, few groups, and low ICCs (rejection rate = .78). Otherwise, Level-2 RMSEA small model rejection rates ranged from .84 to 1. Excluding the most optimal conditions (many groups and large ICCs), Level-2 RMSEA's power in the large model conditions ranged from 0.24 to .62. For large models, Level-2 RMSEA performed best with many groups and large ICCs (rejection rates = .89 to .93). For large models, Level-2 RMSEA performed worst with asymmetric data, few groups, and small ICCs (rejection rate = .24).

Level-2 SRMR had medium power (.4 to .6) in most conditions. For large models, Level-2 SRMR had power greater than .95 with few groups and small ICCs or asymmetric data with few groups and large ICCs. Situations where Level-2 SRMR had high power to reject misfit usually were accompanied by severe Type I error rates. Otherwise, Level-2 SRMR's power was 0.36 to 0.79. The rejection rate of .79 occurred with large models, asymmetric data, many groups, and small ICCs.

Level-2 fit indices' rejection of Level-2 correct model. Level-2 CFI and TLI rejected the correct model much more frequently than did Level-2 RMSEA and SRMR. Level-2 TLI rejected correct Level-2 models much more frequently than Level-2 CFI. Level-2 TLI's rejection rates were .01 to .28. Level-2 TLI performed worst with small models, many groups, and small ICCs (rejection rate = .27 to .28). Level-2 TLI performed best with large models, symmetric data, few groups, and large ICCs (rejection

rate = .01). Level-2 TLI met desired correct model rejection rates (< .05) only with large models, symmetric data, and large ICCs or large models, asymmetric data, and large ICCs. Level-2 CFI's rejection rates ranged from .01 to .22. Level-2 CFI met desired correct model rejection rates (< .05) only with many groups and large ICCs. Level-2 CFI performed best with large models, symmetric data, few groups, and large ICCs (rejection rate = .01). Level-2 CFI performed worst with small models, asymmetric data, few groups, and small ICCs (rejection rate = .22). Level-2 RMSEA met desired rejection rates (< .05) only when fitting large models. Level-2 RMSEA rejected 0% to 1% of large models and 6% to 20% of small models. For small models, Level-2 RMSEA performed best with symmetric data, many groups, and large ICCs (rejection rate = .06). For small models, Level-2 RMSEA performed worst with asymmetric data, few groups, and small ICCs (rejection rate = .20). Level-2 SRMR met desired rejection rates (<.05) in most conditions. With large models, few groups, and small ICCs, Level-2 SRMR had severe high Type I error, yielding rejection rates of .34 with symmetry and .80 with asymmetry. Level-2 SRMR also rejected 9% of large correct models fit to asymmetric data with few groups and large ICCs. Otherwise, Level-2 SRMR's rejection rates were 0 to .01.

Data Asymmetry's Impact on Level-2 Fit Indices

Before describing asymmetry's impact on Level-2 fit indices, figures will be shown to introduce the complex nature of level-specific fit index performance. Data asymmetry's impact generally also depended on the fit index, type of misspecification, model size, number of groups, and ICCs. Figures 28 and 29 contrast Level-2 fit indices'

rejection rates with symmetric and severely asymmetric data for various conditions.

Tables 6 and 7 give Level-2 fit indices' rejection rates.

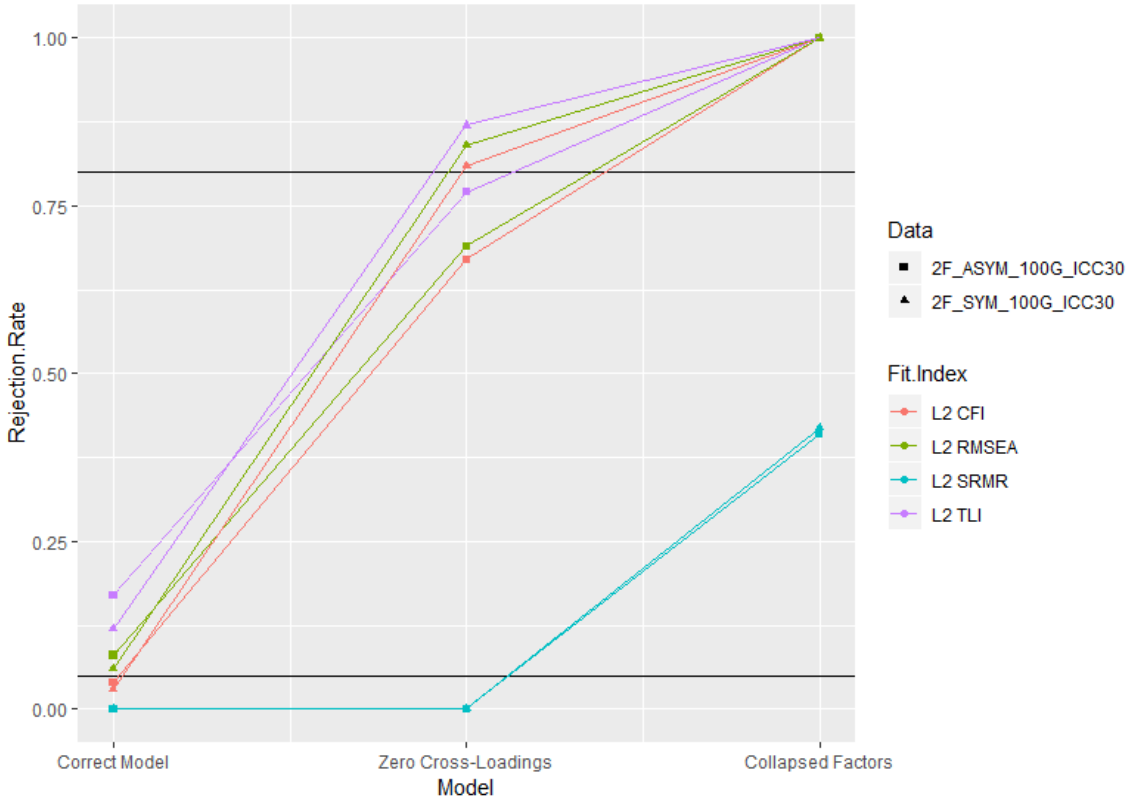

Figure 28. Impact of Data Asymmetry on Level-2 Fit Indices' Rejection Rates with Small Models, Many Groups, and Large ICCs.
Note. This figure has considerable overlap. Level-2 SRMR's rejection rates were zero for correct models and cross-loadings fixed to zero regardless of asymmetry. For collapsed factors, Level-2 CFI, TLI, and RMSEA's rejection rates were 1 regardless of symmetry. For collapsed factors, Level-2 SRMR's rejection rates differed by .01 for symmetric data and asymmetric data.

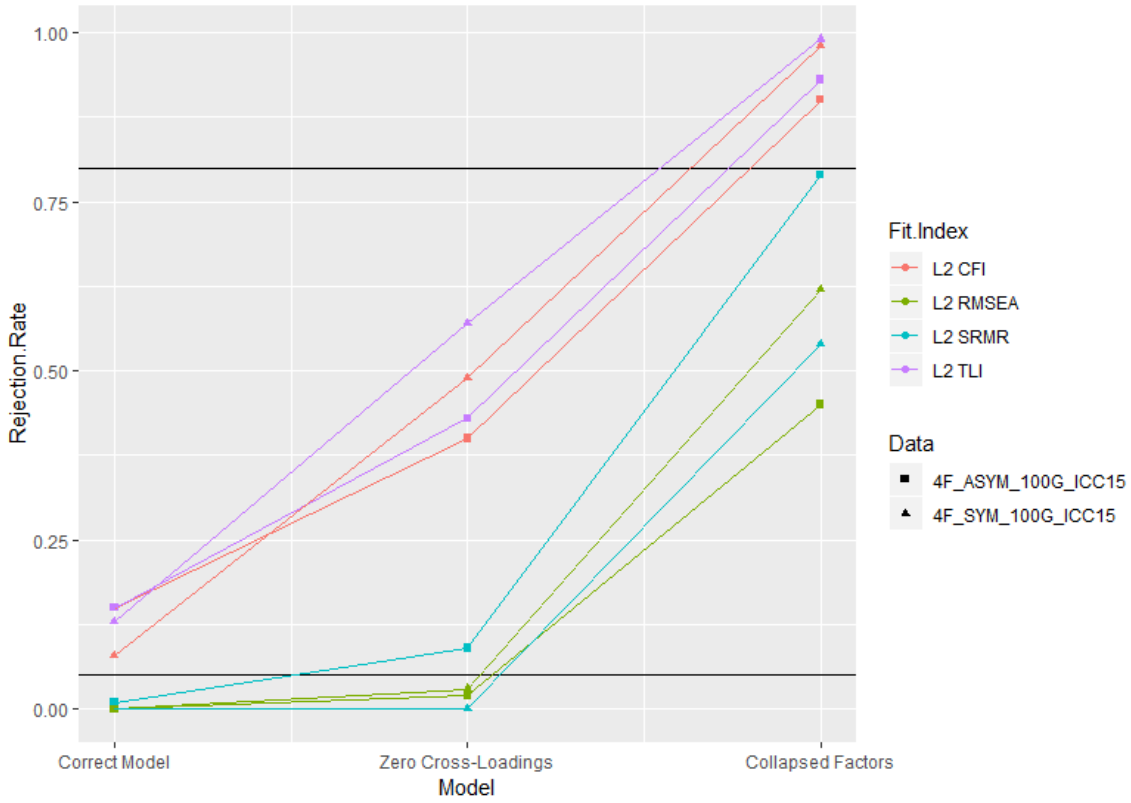

Figure 29. Impact of Data Asymmetry on Level-2 Fit Indices' Rejection Rates with Large Models, Many Groups, and Small ICCs.
Note. This figure has overlap. For correct models, rejection rates were zero for Level-2 RMSEA with asymmetry and for Level-2 RMSEA and SRMR with symmetry. For correct models, rejection rates were .15 for Level-2 CFI and TLI with asymmetry.

Level-2 fit indices' rejection of Level-2 cross-loadings fixed to 0. Level-2 CFI rejection of Level-2 cross-loadings fixed to 0 usually decreased as data asymmetry increased. Level-2 CFI rejection rates were .16 to .81 with symmetry and .20 to .67 with asymmetry. Level-2 CFI performed best with small models, many groups, and large ICCs, yielding rejection rates of .81 with symmetry and .67 with asymmetry. Level-2 CFI performed worst with large models, producing rejection rates of .16 to .49 with symmetry and .2 to .43 with asymmetry. With large models, few groups, and large ICCs, Level-2 CFI rejection rates were .16 with symmetry and .20 with asymmetry. With large models, few groups, and small ICCs, Level-2 CFI was unaffected by asymmetry (rejection rate = .33). With few groups and large ICCs, Level-2 CFI's rejection rates increased as asymmetry increased.

Generally, Level-2 TLI rejection of Level-2 cross-loadings fixed to 0 decreased as data asymmetry increased. Level-2 TLI rejection rates were .20 to .87 with symmetry and .24 to .77 with asymmetry. Level-2 TLI performed best with small models, many groups, and large ICCs, yielding rejection rates of .87 with symmetry and .77 with asymmetry. Level-2 TLI performed worst with large models and few groups, producing rejection rates of .20 to .40 with symmetry and .24 to .35 with asymmetry. Level-2 TLI was unaffected by asymmetry with large models, many groups, and large ICCs (rejection rate = .54). With few groups and large ICCs, Level-2 TLI's rejection rates increased as asymmetry increased.

Generally, Level-2 RMSEA's rejection of Level-2 cross-loadings fixed to 0 decreased as data asymmetry increased. Level-2 RMSEA rejection rates were .03 to .84

with symmetry and .02 to .69 with asymmetry. Level-2 RMSEA performed worst with large models. For large models, Level-2 RMSEA rejection rates were .03 to .11 with symmetry and .02 to .06 with asymmetry. For small models, Level-2 RMSEA rejection rates were .49 to .84 with symmetry and .39 to .69 with asymmetry. Level-2 RMSEA performed best with small models, many groups, and large ICCs. In these situations, Level-2 RMSEA rejection rates were .84 with symmetry and .69 with asymmetry. With small models, few groups, and large ICCs, Level-2 RMSEA's rejection rates increased .01 as asymmetry increased.

Generally, Level-2 SRMR's rejection rates of Level-2 cross-loadings fixed to 0 increased as data asymmetry increased. Level-2 SRMR rejection rates were 0 to .67 with symmetry and 0 to .89 with asymmetry. Level-2 SRMR performed worst with many groups, producing rejection rates of 0 to .02 with symmetry and 0 to .09 with asymmetry. Level-2 SRMR performed best with large models, few groups, and small ICCs. In this situation, Level-2 SRMR rejection rates were .67 with symmetry and .89 with asymmetry. With small models, many groups, and low ICCs Level-2 SRMR's rejection rates decreased as asymmetry increased. In this situation, Level-2 SRMR rejection rates were .05 with symmetry and .11 with asymmetry.

Level-2 fit indices' rejection of Level-2 collapsed factors. Generally, Level-2 CFI rejection rates of Level-2 collapsed factors *decreased* as data asymmetry increased. Level-2 CFI rejection rates were .81 to 1 with symmetry and .75 to 1 with asymmetry. Level-2 CFI performed worst with large models, few groups, and small ICCs, yielding rejection rates of .81 with symmetric data and .75 with asymmetric data. Level-2 CFI

performed best with small models, many groups, and large ICCs. In this situation, Level-2 CFI rejected 100% of collapsed Level-2 models regardless of asymmetry. Level-2 CFI was most affected by asymmetry with large models, many groups, and small ICCs. In this situation, Level-2 CFI rejection rates were .98 with symmetric data and .90 with asymmetric data. With large models, few groups and large ICCs, Level-2 CFI rejection rates increased as asymmetry increased.

Generally, Level-2 TLI rejection rates of Level-2 collapsed factors *decreased* as data asymmetry increased. Level-2 TLI rejection rates were .83 to 1 with symmetry and .77 to 1 with asymmetry. Level-2 TLI performed worst with large models and few groups, yielding rejection rates of .83 to .85 with symmetric data and .77 to .84 with asymmetric data. Level-2 TLI performed best with small models, many groups, and large ICCs. In this situation, Level-2 TLI rejected 100% of collapsed Level-2 models regardless of asymmetry. With large models, few groups and large ICCs, Level-2 TLI rejection rates increased as asymmetry increased.

Generally, Level-2 RMSEA rejection rates of Level-2 collapsed factors *decreased* as data asymmetry increased. Level-2 RMSEA rejection rates were .41 to 1 with symmetry and .24 to 1 with asymmetry. Level-2 RMSEA performed worst with large models and few groups. In these situations, Level-2 RMSEA rejection rates were .41 to .48 with symmetric data and .24 to .50 with asymmetric data. Level-2 RMSEA performed best with small models, many groups, and large ICCs. In this situation, Level-2 RMSEA rejection rates were 1 regardless of asymmetry. Level-2 RMSEA performed much worse with large models than small models. For large models, Level-2 RMSEA rejection rates

were .41 to .93 with symmetric data and .24 to .89 with asymmetric data. For small models, Level-2 RMSEA rejection rates were .84 to 1 with symmetric data and .78 to 1 with asymmetric data. With large models, few groups and large ICCs, Level-2 RMSEA rejection rates increased as asymmetry increased.

Generally, Level-2 SRMR rejection rates of Level-2 collapsed factors *increased* as data asymmetry increased. Level-2 SRMR rejection rates were .36 to .96 with symmetry and .41 to .91 with asymmetry. Level-2 SRMR performed worst with many groups and large ICCs, yielding rejection rates of .36 to .42 with symmetry and .41 to .53 with asymmetry. Level-2 SRMR performed best with large models and few groups, producing rejection rates of .87 to .96 with symmetry and .95 to .99 with asymmetry. Level-2 SRMR rejection decreased with small models, many groups, and large ICCs.

Level-2 fit indices' rejection of correct Level-2 model. Level-2 CFI rejection rates of correct model generally increased as asymmetry increased. Level-2 CFI rejection rates were .03 to .18 with symmetry and .04 to .22 with asymmetry. Level-2 CFI performed worst with few groups and small ICCs, producing correct model rejection rates of .18 to .20 with symmetric data and .16 to .22 with asymmetric data. Level-2 CFI performed best with many groups and large ICCs, yielding correct model rejection rates of .02 to .03 with symmetry and .03 to .04 with asymmetry. Level-2 CFI was most affected by asymmetry with small models, few groups, and large ICCs. In these situations, Level-2 CFI rejection rates were .07 with symmetry and .16 with asymmetry (a .09 increase). With large models, few groups, and small ICCs, Level-2 CFI rejection of correct models decreased as asymmetry increased.

Level-2 TLI rejection of the correct model usually increased as asymmetry increased. Level-2 TLI rejection rates were .01 to .13 with symmetry and .07 to .18 with asymmetry. Level-2 TLI performed worst with small models, few groups, and small ICCs, producing correct model rejection rates of .26 regardless of asymmetry. Level-2 TLI performed best with large models, many groups, and large ICCs, yielding rejection rates of .04 regardless of asymmetry. Level-2 TLI correct model rejection was most affected by asymmetry with large models, few groups, and large ICCs. In these situations, Level-2 TLI rejection rates were .01 with symmetry and .07 with asymmetry. Level-2 TLI rejection rates decreased as asymmetry increased for large models, few groups, and small ICCs and for small models, many groups, and small ICCs. In these situations, Level-2 TLI rejection rates were .22 to .28 with symmetry and .20 to .27 with asymmetry (decreases of .01 to .02).

Asymmetry affected Level-2 RMSEA rejection of correct Level-2 models differently based on model size. Level-2 RMSEA rejection of large correct models was unaffected by asymmetry. For large models, Level-2 RMSEA rejection rates were 0 to .01 with symmetry and 0 to .01 with asymmetry. Level-2 RMSEA's rejection of small correct models usually increased as asymmetry increased. Level-2 RMSEA rejection rates were .06 to .16 with symmetry and .08 to .20 with asymmetry. Asymmetry most affected Level-2 RMSEA with small models, few groups, and large ICCs. In this situation, Level-2 RMSEA rejection rates were .07 with symmetry and .15 with asymmetry (increase of .08). For small models, Level-2 RMSEA performed best with many groups and large ICCs, yielding rejection rates of .06 with symmetry and .08 with

asymmetry. With many groups and small ICCs, Level-2 RMSEA rejection of small correct models decreased as asymmetry increased. In this situation, Level-2 RMSEA rejection rates were .12 with symmetry and .10 with asymmetry.

Asymmetry differentially affected Level-2 SRMR rejection of correct models based on model size. Level-2 SRMR usually never rejected small correct models regardless of asymmetry. However, with small models, few groups, and small ICCs, Level-2 SRMR rejection rates were .01 with symmetry and .09 with asymmetry. Level-2 SRMR rejection rates usually increased as asymmetry increased for most large correct models. Level-2 SRMR rejection rates increased between .01 and .46, yielding rejection rates of 0 to .34 with symmetry and .09 to .80 with asymmetry. Level-2 SRMR performed worst with large models, few groups, and small ICCs, yielding rejection rates of .34 with symmetry and .80 with asymmetry. Otherwise, Level-2 SRMR rejection rates were 0 to .01 with symmetry and 0 to .09 with asymmetry. Level-2 SRMR performed best with small models, usually yielding rejection rates of 0 regardless of asymmetry.

Model Size's Impact on Level-2 Fit Indices

Before describing the impact of model size on Level-2 fit indices' rejection rates, some figures will be shown. These figures introduce the complex nature of level-specific fit index performance. Model size's impact generally also depended on the specific fit index, type of misspecification, asymmetry, number of groups, and ICCs. Figures 30 through 32 contrast Level-2 fit indices' rejection rates with small and large models for various combinations of conditions. Recall that Tables 6 and 7 give Level-2 fit indices' rejection rates.

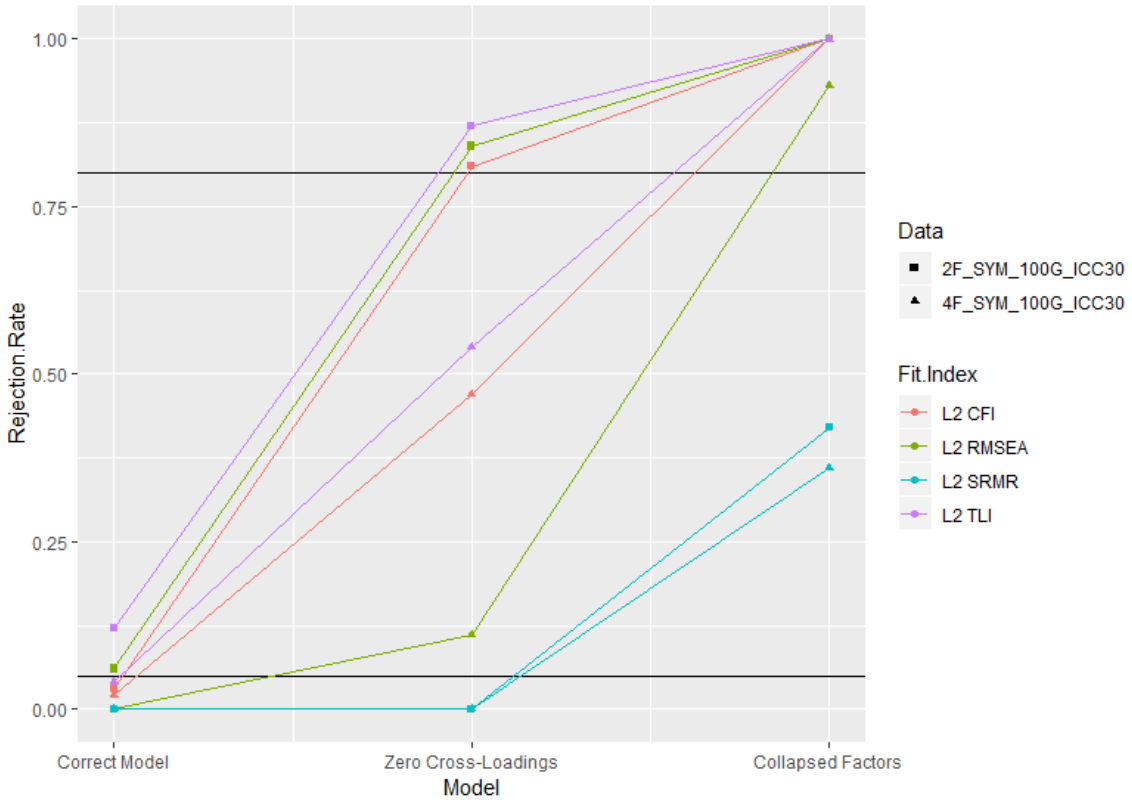

Figure 30. Impact of Model Size on Level-2 Fit Indices' Rejection Rates with Symmetric Data, Many Groups, and Large ICCs.

Note. This figure has considerable overlap. For correct models, rejection rates were zero for Level-2 SRMR with small models and for Level-2 RMSEA and SRMR with large models. For zero cross-loadings, Level-2 SRMR rejection rates were zero regardless of model size. For collapsed factors, rejection rates were 1 for Level-1 CFI, TLI, and RMSEA with small models and for Level-1 CFI and TLI with large models.

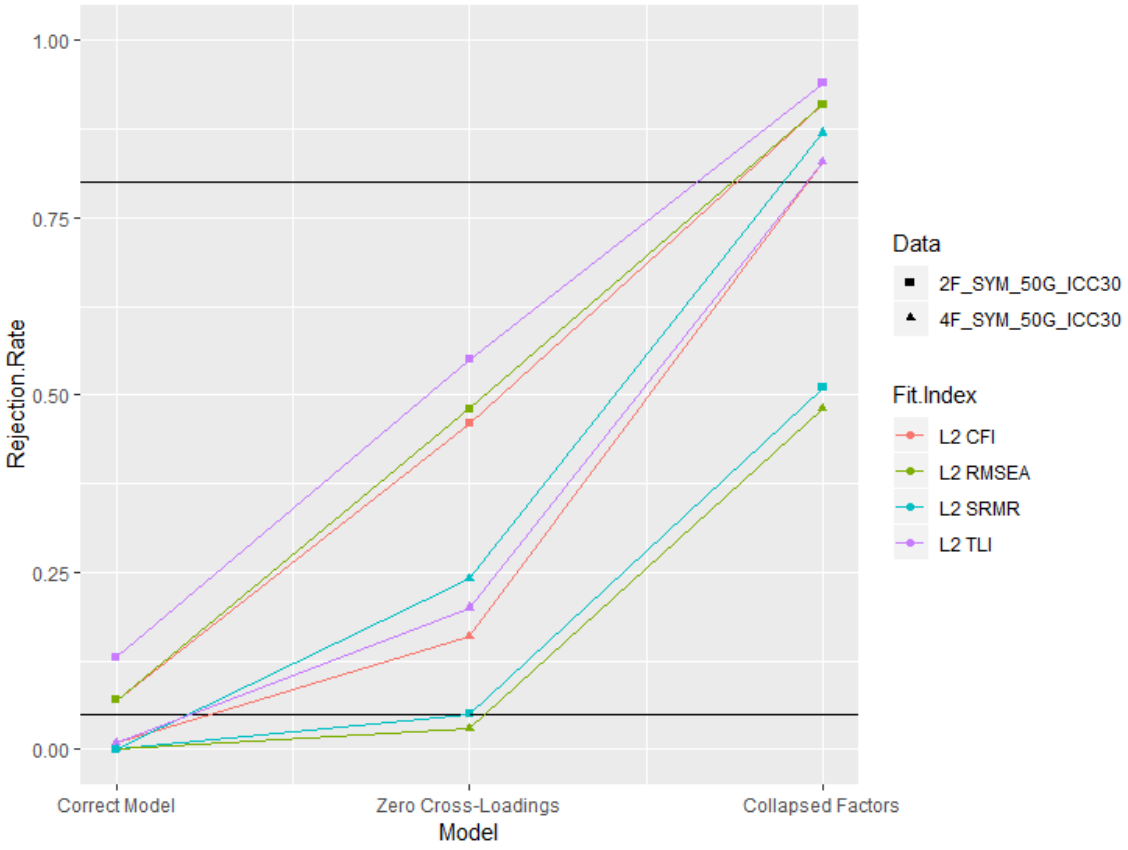

Figure 31. Impact of Model Size on Level-2 Fit Indices' Rejection Rates with Symmetric Data, Few Groups, and Large ICCs.

Note. This figure has overlap. For correct models, Level-2 CFI and TLI had identical rejection rates with large models. For correct models, rejection rates were zero for Level-2 SRMR with small models and for Level-2 RMSEA and SRMR with large models. For collapsed factors, Level-2 CFI and TLI had identical rejection rates for large models. For collapsed factors, Level-2 CFI and RMSEA had identical rejection rates for small models.

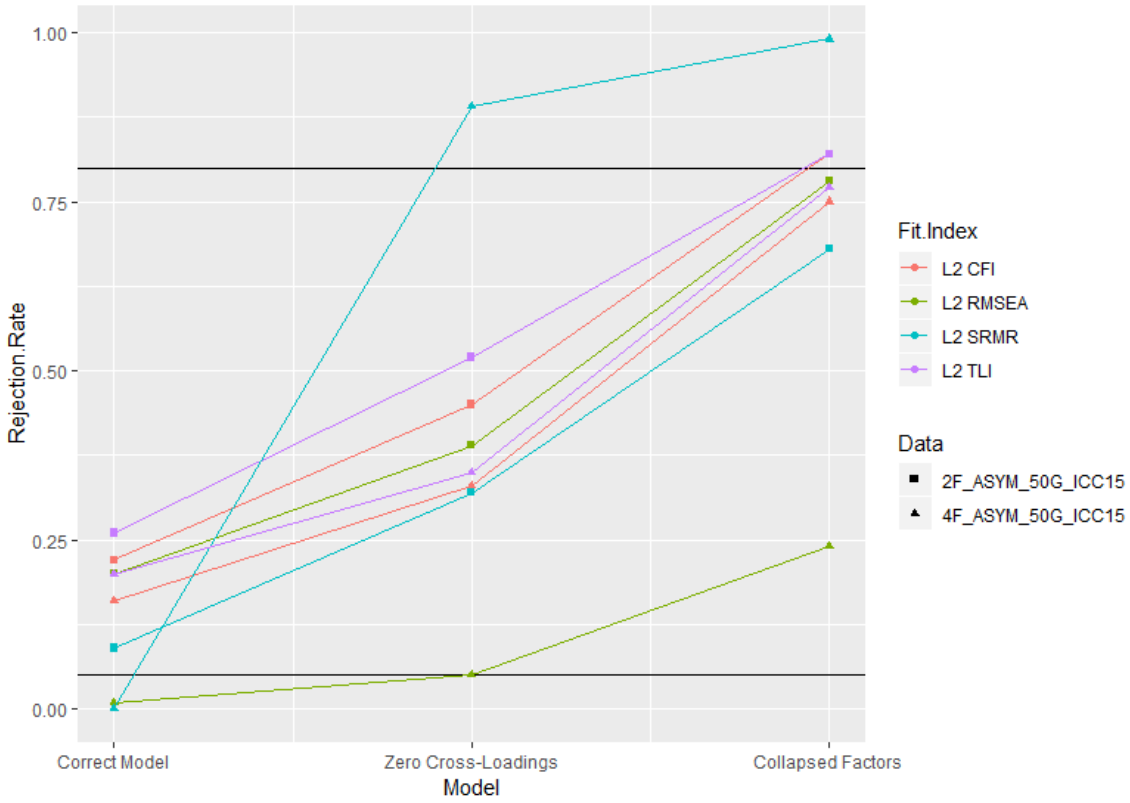

Figure 32. Impact of Model Size on Level-2 Fit Indices' Rejection Rates with Asymmetric Data, Few Groups, and Small ICCs.
Note. This figure has overlap. For correct models, rejection rates were identical for Level-2 RMSEA with small models and Level-2 TLI with large models. For zero cross-loadings and collapsed factors, Level-2 CFI and TLI's rejection rates only differed by .02 for large models. For collapsed factors, Level-2 CFI and TLI had identical rejection rates with small models.

Level-2 fit indices' rejection of Level-2 cross-loadings fixed to 0. Level-2 CFI rejection of Level-2 crossloadings fixed to 0 always decreased as model size increased. Level-2 CFI rejection rates were .45 to .81 with small models and .16 to .49 with large models. Level-2 CFI performed worst with asymmetric data and few groups, producing rejection rates of .45 to .50 with small models and .20 to .33 with large models. Level-2 CFI performed best with symmetric data, many groups, and large ICCs. In this situation,

Level-2 CFI rejection rates were .81 with small models and .47 with large models (the largest increase of .34). Model size most affected Level-2 CFI with symmetric data, many groups, and large ICCs. Model size least affected with small models, few groups, and small ICCs. In this situation, Level-2 CFI rejection rates were .45 with small models and .33 with large models (the smallest decrease of .12).

Level-2 TLI rejection of Level-2 crossloadings fixed to 0 always decreased as model size increased. Level-2 TLI rejection rates were .52 to .87 with small models and .20 to .54 with large models. Level-2 TLI performed best with symmetric data and many groups. These situations yielded Level-2 TLI rejection rates of .76 to .87 with small models and .54 to .57 with large models. Level-2 TLI performed worst with few groups and large ICCs, yielding rejection rates of .55 to .59 with small models and .20 to .24 with large models.

Level-2 RMSEA rejection of Level-2 crossloadings fixed to 0 always decreased as model size increased. Level-2 RMSEA rejection rates were .34 to .84 with small models and .02 to .11 with large models. Model size most affected Level-2 RMSEA with symmetric data, many groups, and large ICCs, yielding rejection rates of .84 with small models and .11 with large models (the largest decrease of .73). Level-2 RMSEA performed worst with asymmetry, few groups, and large ICCs, yielding rejection rates of .39 with small models and .05 with large models.

Level-2 SRMR rejection rates of Level-2 cross-loadings fixed to 0 usually increased as model size increased. Level-2 SRMR rejection rates were .05 to .32 with small models and .09 to .89 with large models. Level-2 SRMR performed best with

severe asymmetry, few groups, and small ICCs, producing rejection rates of .32 with small models and .89 with large models. Level-2 SRMR performed worst with many groups and large ICCs, yielding rejection rates of 0 regardless of model size. With symmetric data, many groups, and small ICCs, Level-2 SRMR rejection rates decreased from .02 with small models to 0 with large models.

Level-2 fit indices' rejection of Level-2 collapsed factors. Level-2 CFI rejection of collapsed Level-2 factors usually *decreased* as model size increased. Level-2 CFI rejection rates were .82 to 1 with small models and .75 to 1 with large models. Level-2 CFI performed worst with asymmetry, few groups, and small ICCs, yielding rejection rates of .82 with small models and .75 with large models. Otherwise, Level-2 CFI rejection rates were .87 to 1 with small models and .81 to 1 with large models. Level-2 CFI performed best with many groups and large ICCs, yielding rejection rates of 1 regardless of model size.

Level-2 TLI rejection of collapsed Level-2 factors usually *decreased* as model size increased. Level-2 TLI rejection rates were .82 to 1 with small models and .77 to 1 with large models. Level-2 TLI performed worst with asymmetry, few groups, and small ICCs, yielding rejection rates of .82 with small models and .77 with large models. Otherwise, Level-2 TLI rejection rates were .88 to 1 with small models and .83 to 1 with large models. Level-2 TLI performed best with many groups and large ICCs, yielding rejection rates of 1 regardless of model size.

Level-2 RMSEA rejection of collapsed Level-2 factors always *decreased* as model size increased. Level-2 RMSEA rejection rates were .78 to 1 with small models

and .24 to .93 with large models. Level-2 RMSEA performed best with many groups and large ICCs, yielding rejection rates of 1 with small models and .89 to .93 with large models. Level-2 RMSEA performed worst with asymmetric data, few groups, and small ICCs, yielding rejection rates of .78 with small models and .24 with large models.

Level-2 SRMR rejection rates of Level-2 collapsed factors usually *increased* as model size increased. Level-2 SRMR rejection rates were .39 to .68 for small models and .53 to .99 for large models. Level-2 SRMR performed worst with many groups and large ICCs, yielding rejection rates of .39 with small models and .54 with large models. Level-2 SRMR performed best with few groups and asymmetry, yielding rejection rates of .6 to .68 with small models and .95 to .99 with large models. With approximate symmetry, many groups, and large ICCs, Level-2 SRMR rejection rates decreased as model size increased. In this situation, Level-2 SRMR rejection rates were .42 with small models and .36 with large models.

Level-2 fit indices' rejection of Level-2 correct models. Level-2 CFI rejection of correct Level-2 models usually decreased as model size increased. Level-2 CFI rejection rates were .03 to .22 for small models and .01 to .16 for large correct models. Level-2 CFI performed worst with severe asymmetry, few groups, and low ICCs. In these situations, Level-2 CFI rejected 22% of small models and 16 % of large models. Level-2 CFI performed best with symmetric data, many groups, and large ICCs. In these situations, Level-2 CFI rejected 3% of small models and 2% of large models. Level-2 CFI was unaffected by model size with asymmetry, many groups, and small ICCs (rejection rate = .15). With symmetric data, few groups, and small ICCs, Level-2 CFI

rejection rates increased as model size increased. In these situations, Level-2 CFI rejected 18% of small correct models and 20% of large correct models.

Level-2 TLI rejection of correct Level-2 models always decreased as model size increased. Level-2 TLI rejection rates were .12 to .28 with small models and .01 to .22 with large models. Level-2 TLI performed worst with small ICCs, yielding rejection rates of .26 to .28 with small models and .13 to .22 with large models. Level-2 TLI performed best with symmetric data and large ICCs, yielding rejection rates of .12 to .13 with small models and .01 to .04 with large models.

Level-2 RMSEA rejection of correct Level-2 models always decreased as model size increased. Level-2 RMSEA rejection rates were .06 to .20 for small correct models and 0 to .01 for large correct models. For small models, Level-2 RMSEA performed best with symmetric data and large ICCs, yielding rejection rates of .06 to .07. For small models, Level-2 RMSEA performed worst with asymmetry, few groups, and small ICCs, yielding rejection rates of .20.

Level-2 SRMR rejection of Level-2 correct models either increased as model size increased or remained 0 regardless of model size. Level-2 SRMR rejection rates were .01 to .09 for small correct models and .01 to .80 for large correct models. For small models, Level-2 SRMR performed worst with asymmetry, few groups, and small ICCs (rejection rate = .09). Otherwise, Level-2 SRMR rejected 0% to 1% of small correct models. For large models, Level-2 SRMR performed worst with few groups and small ICCs. In these situations, Level-2 SRMR rejection rates were .34 with symmetry and .80 with asymmetry. Otherwise, Level-2 SRMR rejection rates were 0 to .08 for large correct

models. Excluding the large model condition with asymmetric data, few groups, and large ICCs (rejection rate = .08), Level-2 SRMR rejection rates for correct models were 0 to .01. Level-2 SRMR never rejected correct Level-2 models regardless of model size with many groups. The only exception was with large models, asymmetry, many groups, and small ICCs (rejection rate = .01).

Impact of Number of Groups on Level-2 Fit Indices

Before describing the impact of number of groups on Level-2 fit indices' rejection rates, some figures will be shown. These figures introduce the complex nature of level-specific fit index performance. Number of groups' impact generally also depended on the specific fit index, type of misspecification, asymmetry, model size, and ICCs. Figures 33 through 35 contrast Level-2 fit indices' rejection rates with few and many groups for combinations of conditions. Tables 6 and 7 give Level-2 fit indices' rejection rates.

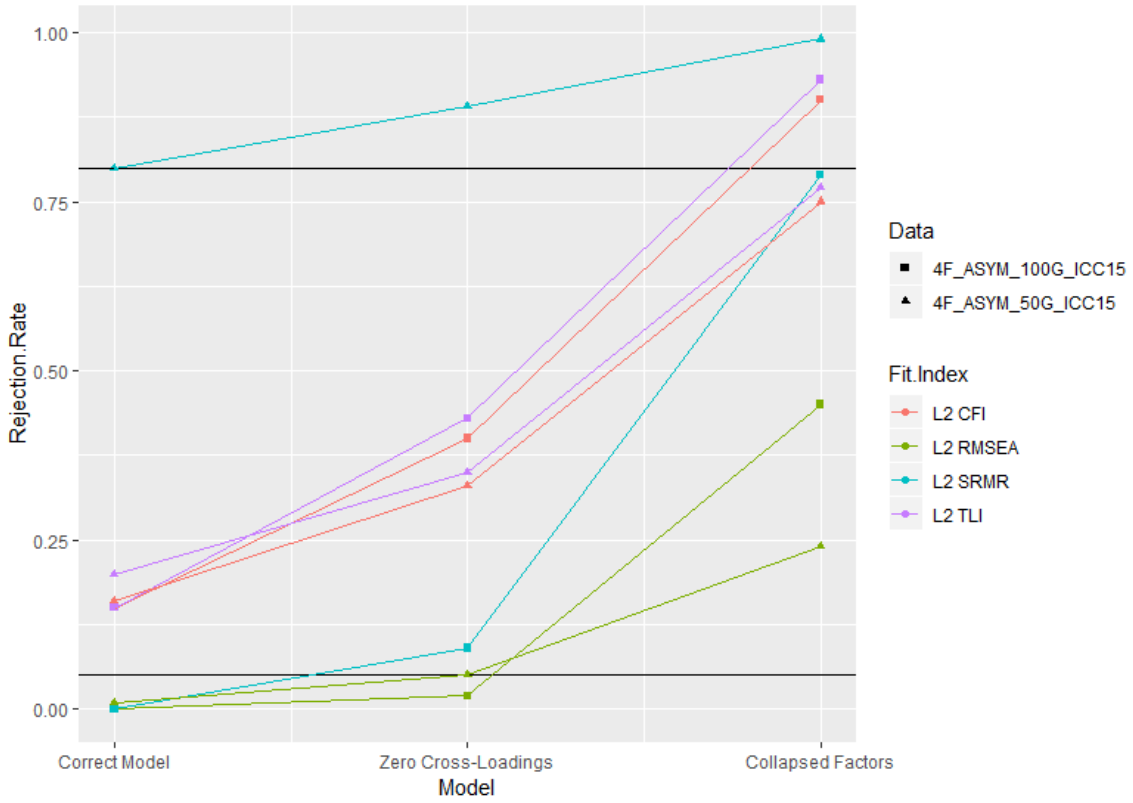

Figure 33. Impact of Number of Groups on Level-2 Fit Indices' Rejection of Large Models, Asymmetric Data, and Small ICCs.
Note. This figure has considerable overlap. For correct models, Level-2 CFI and TLI had identical rejection rates with many groups; they differed .01 from Level-2 CFI's rejection rate with few groups. For correct models, rejection rates were identical for Level-2 RMSEA with few groups and Level-2 SRMR with many groups; they differed by .01 from Level-2 RMSEA's rejection rate for correct models with many groups.

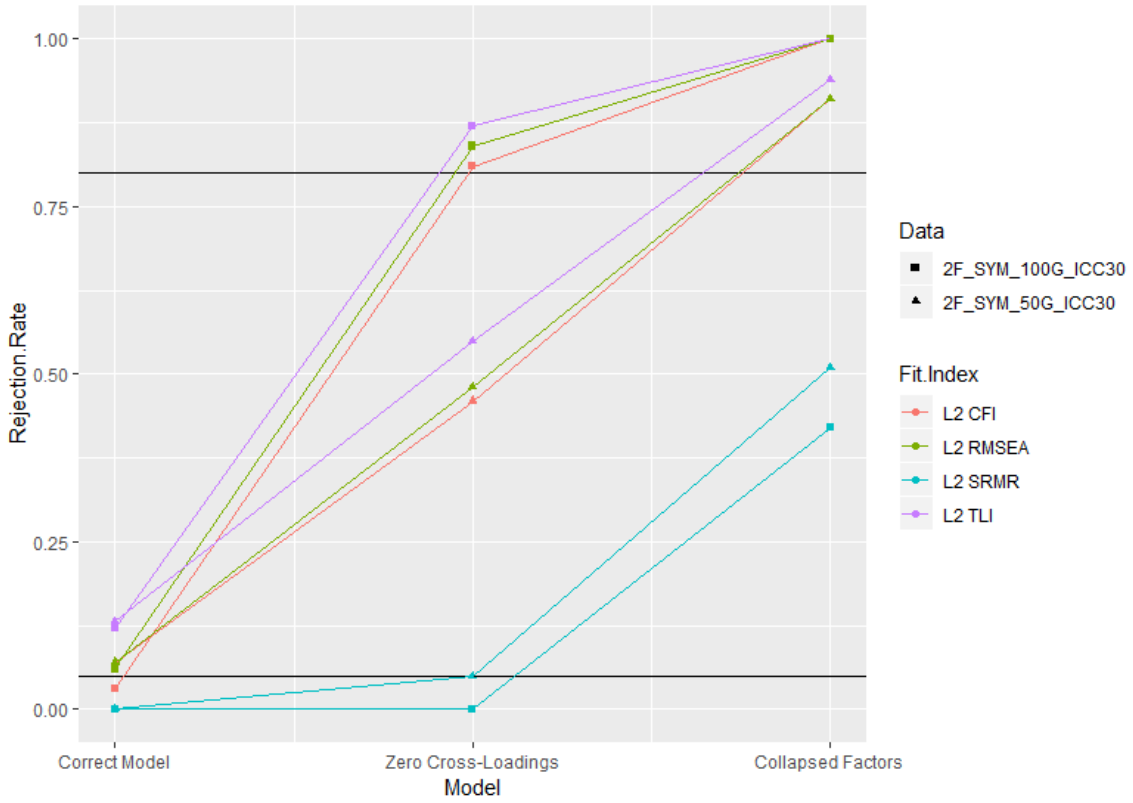

Figure 34. Impact of Number of Groups on Level-2 Fit Indices' Rejection of Small Models, Symmetric Data, and Large ICCs.
Note. This figure has overlap. For correct models, Level-2 CFI and RMSEA had identical rejection rates with few groups; they differed .01 from Level-2 RMSEA's rejection rate with many groups. For correct models, Level-2 SRMR rejection rates were zero for few and many groups. For correct models, Level-2 TLI's rejection rate differed by .01 for few and many groups. For collapsed factors, Level-2 CFI and RMSEA had identical rejection rates with few groups. For collapsed factors, Level-2 CFI, TLI, and RMSEA had identical rejection rates with many groups.

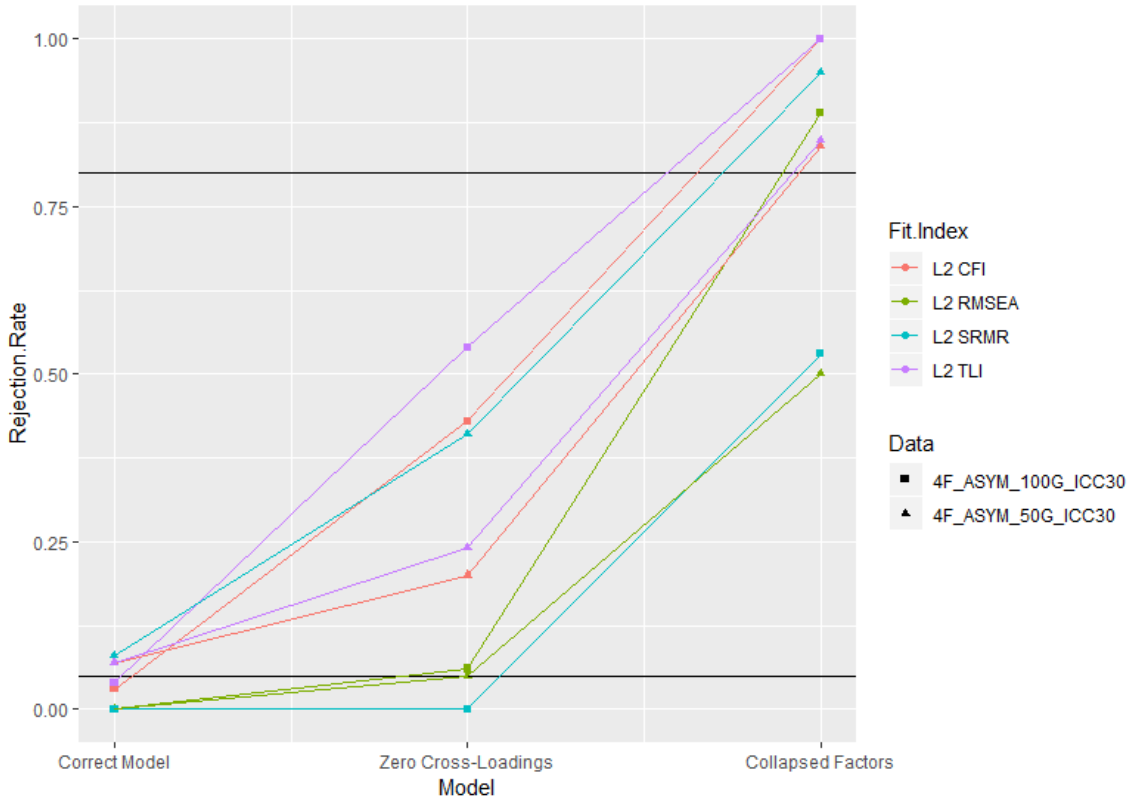

Figure 35. Impact of Number of Groups on Level-2 Fit Indices' Rejection of Large Models, Asymmetric Data, and Large ICCs.
Note. This figure has overlap. For correct models, rejection rates were zero for Level-2 RMSEA regardless of number of groups and for Level-2 SRMR with many groups. For correct models, Level-2 CFI and TLI had identical rejection rates with few groups. For collapsed factors, Level-2 CFI and TLI had identical rejection rates with many groups. For collapsed factors, Level-2 CFI and TLI's rejection rates differed by .01 with few groups.

Level-2 fit indices' rejection of Level-2 cross-loadings fixed to 0. Level-2 CFI rejection of Level-2 cross-loadings fixed to 0 always increased as number of groups increased. Level-2 CFI rejection rates were .2 to .52 with few groups and .4 to .81 with many groups. Level-2 CFI performed worst with large models and large ICCs. These situations produced rejection rates of .16 to .20 with few groups and .43 to .47 with many groups. With few groups, Level-2 CFI performed best with small models, symmetric data, and small ICCs (rejection rate = .52). With many groups, Level-2 CFI performed best with small models, symmetric data, and large ICCs (rejection rate = .81). Level-2 CFI Level-2 CFI rejection rates increased the most (.35) as number of groups increased with small models, symmetry, and large ICCs. These conditions produced rejection rates of .46 with few groups and .81 with many groups. Level-2 CFI rejection rates increased the least (.07) with large models, severe asymmetry, and small ICCs. These conditions produced rejection rates of .33 with few groups and .40 with many groups.

Level-2 TLI rejection of Level-2 cross-loadings fixed to 0 always increased as number of groups increased. Level-2 TLI rejection were .2 to .59 with few groups and .43 to .87 with many groups. With few groups, Level-2 TLI performed worst with large models and large ICCs. These conditions yielded rejection rates of .20 to .24. With many groups, Level-2 TLI performed worst with large models, asymmetric data, and small ICCs. This situation produced rejection rates of .43. Level-2 TLI performed best with small models and large ICCs. These conditions produced rejection rates of .55 to .59 with few groups and .77 to .87 with many groups. Level-2 TLI performance improved most as number of groups increased with symmetric data and large ICCs (increases between .32

and .34). Level-2 TLI performance improved least as number of groups increased with asymmetric data and small ICCs (increase of .08).

Level-2 RMSEA rejection of Level-2 cross-loadings fixed to 0 usually increased as number of groups increased. Level-2 RMSEA rejection rates were .05 to .49 with few groups and .06 to .84 with many groups. Level-2 RMSEA performed worst with large models. These conditions produced rejection rates of .04 to .05 with few groups and .02 to .11 with many groups. Level-2 RMSEA performed best with small models, symmetry, and large ICCs. These conditions produced rejection rates of .48 with few groups and .84 with many groups (the largest increase of .36). However, with few groups, Level-2 RMSEA rejection rates were slightly higher (rejection rate = .49) with small models and either asymmetry and large ICCs or symmetry and small ICCs. Level-2 RMSEA rejection rates decreased with large models and low ICCs. These conditions produced rejection rates of .04 to .05 with few groups and .02 to .03 with many groups.

Level-2 SRMR rejection of Level-2 cross-loadings fixed to 0 always decreased as number of groups increased. Level-2 SRMR rejection rates were.05 to .89 with few groups and 0 to .09 with many groups. Level-2 SRMR performed worst with small models, symmetry, and large ICCs. These conditions produced rejection rates of .05 with few groups and 0 with many groups. Level-2 SRMR performed best with large models, severe asymmetry, and small ICCs. These conditions produced rejection rates of .89 with few groups and .09 with many groups.

Level-2 fit indices' rejection of collapsed Level-2 factors. Level-2 CFI's rejection of Level-2 collapsed factors always increased as number of groups increased.

Level-2 CFI rejection rates were .75 to .91 with few groups and .9 to 1 with many groups. Level-2 CFI performed worst with large models, severe asymmetry, and low ICCs. These conditions yielded rejection rates of .75 with few groups and .9 with many groups. Level-2 CFI performed best with small models, symmetry, and large ICCs. These conditions yielded rejection rates of .91 with few groups and 1 with many groups.

Level-2 TLI rejection of Level-2 collapsed factors always increased as number of groups increased. Level-2 TLI rejection rates were .77 to .94 with few groups and .93 to 1 with many groups. Level-2 TLI performed worst with large models, severe asymmetry, and low ICCs. These conditions yielded rejection rates of .77 with few groups and .93 with many groups. Level-2 TLI performed best with small models, symmetry, and large ICCs. These conditions yielded rejection rates of .94 with few groups and 1 with many groups.

Level-2 RMSEA rejection of Level-2 collapsed factors always increased as number of groups increased. Level-2 RMSEA rejection rates were .24 to 1 with few groups and .45 to 1 with many groups. Level-2 RMSEA performed the worst with large models, severe asymmetry, and low ICCs. These situations yielded rejection rates of .24 with few groups and .45 with many groups. Level-2 RMSEA performed best with small models, symmetric data, and large ICCs. These situations produced rejection rates of .91 with few groups and 1 with many groups. Level-2 RMSEA's largest increase in rejection rates occurred with large models, severe asymmetry, and large ICCs. Level-2 RMSEA's rejection rate improved from .5 with few groups to .89 with many groups.

Level-2 SRMR rejection of Level-2 collapsed factors always decreased as number of groups increased. Level-2 SRMR rejection rates were .51 to .99 with few groups and .39 to .79 with many groups. Level-2 SRMR performed worst with small models, symmetry, and large ICCs. These conditions produced rejection rates of .51 with few groups and .42 with many groups. Level-2 SRMR usually had high rejection rates for correct and incorrect models with large models, severe asymmetry, few groups, and low ICCs, yielding rejection rates of .99 with few groups and .79 with many groups. Here, high power with Level-2 SRMR usually was accompanied by very high levels of Type I error. Number of groups most affected Level-2 SRMR with large models, symmetric data, and large ICC, yielding rejection rates of .87 with few groups and .36 with many groups. Number of groups least affected Level-2 SRMR with small models, symmetry, and large ICCs. These conditions produced rejection rates of .51 with few groups and .42 with many groups (a decrease of .09).

Level-2 fit indices' rejection of correct Level-2 models. Level-2 CFI rejection of correct Level-2 models usually decreased as number of groups increased. Level-2 CFI rejection rates were .01 to .22 for few groups and .02 to .15 for many groups. Level-2 CFI performed best with large models, symmetric data, and large ICCs. These situations produced rejection rates of .01 with few groups and .02 with many groups. Level-2 CFI performed worst with low ICCs, yielding rejection rates of .16 to .22 with few groups and .08 to .15 with many groups.

Level-2 TLI rejection of correct Level-2 models usually decreased as number of groups increased. Level-2 TLI rejection rates were .13 to .22 with few groups and .12 to

.15 with many groups. Level-2 TLI performed best with large models, symmetric data, and large ICCs. These situations yielded rejection rates of .01 with few groups and .04 with many groups. Level-2 TLI performed worst with many groups and low ICCs. These situations produced rejection rates of .26 with few groups and .28 with many groups.

Level-2 RMSEA rejection of correct Level-2 models usually decreased as number of groups increased. Level-2 RMSEA rejection rates were .01 to .20 with few groups and 0 to .12 with many groups. For large models, Level-2 RMSEA rejection rates were 0 to .01 with few groups and 0 with many groups. Increasing number of groups most improved Level-2 RMSEA performance with small model, asymmetric data, and small ICCs. These situations produced rejection rates of .20 with few groups and .10 with many groups. Increasing number of groups least improved Level-2 RMSEA performance with large models (decrease of 0 to .01). For large correct models, Level-2 RMSEA rejection rates were 0 to .01 with few groups and always 0 with many groups.

Level-2 SRMR rejection rates of correct models decreased as number of groups increased or remained at zero regardless of number of groups. Level-2 SRMR rejection rates were 0 to .80 with few groups and 0 to .01 with many groups. The largest decreases in Level-2 SRMR rejection of correct models occurred with small ICCs and large models. With few groups, Level-2 SRMR rejected 34% of large correct models fit to symmetric data with low ICCs and 80% of large correct models fit to severely asymmetric data with low ICCs. Otherwise, Level-2 SRMR rejection rates were 0 to .09 with few groups. Generally, Level-2 SRMR never rejected correct models with large ICCs regardless of

number of groups. However, for large models, large ICCs, and severe asymmetry, Level-2 SRMR rejection rates were .08 with few groups and 0 with many groups.

ICCs' Impact on Level-2 Fit Indices

Before describing the impact of ICCs on Level-2 fit indices' rejection rates, some figures will be shown. These figures introduce the complex nature of level-specific fit index performance. ICCs' impact generally also depended on the specific fit index, type of misspecification, asymmetry, model size, and number of groups. Figures 36 through 38 contrast Level-2 fit indices' rejection rates with small and large ICCs for various combinations of conditions. Tables 6 and 7 give all rejection rates for Level-2 fit indices.

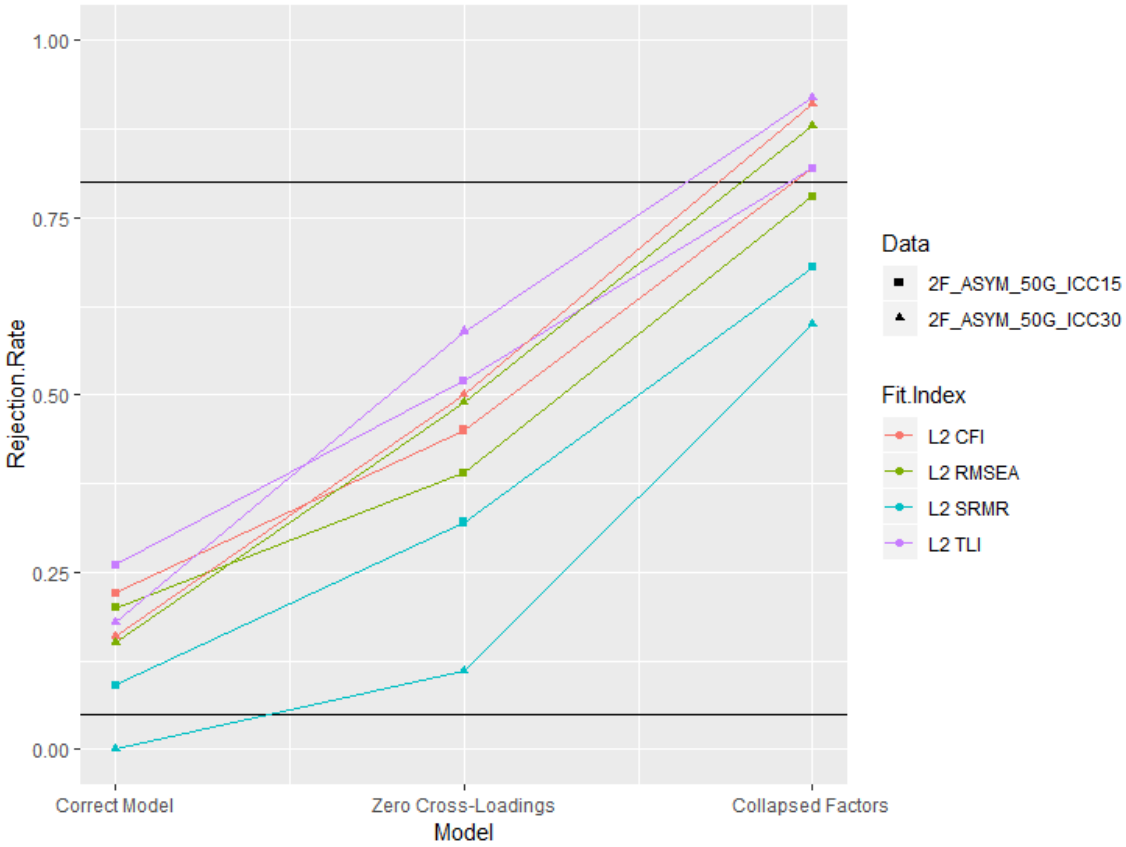

Figure 36. Impact of ICCs on Level-2 Fit Indices' Rejection Rates with Small Models, Asymmetric Data, and Few Groups.
Note. This figure has overlap. For collapsed factors, Level-2 CFI and TLI had identical rejection rates with small ICCs.

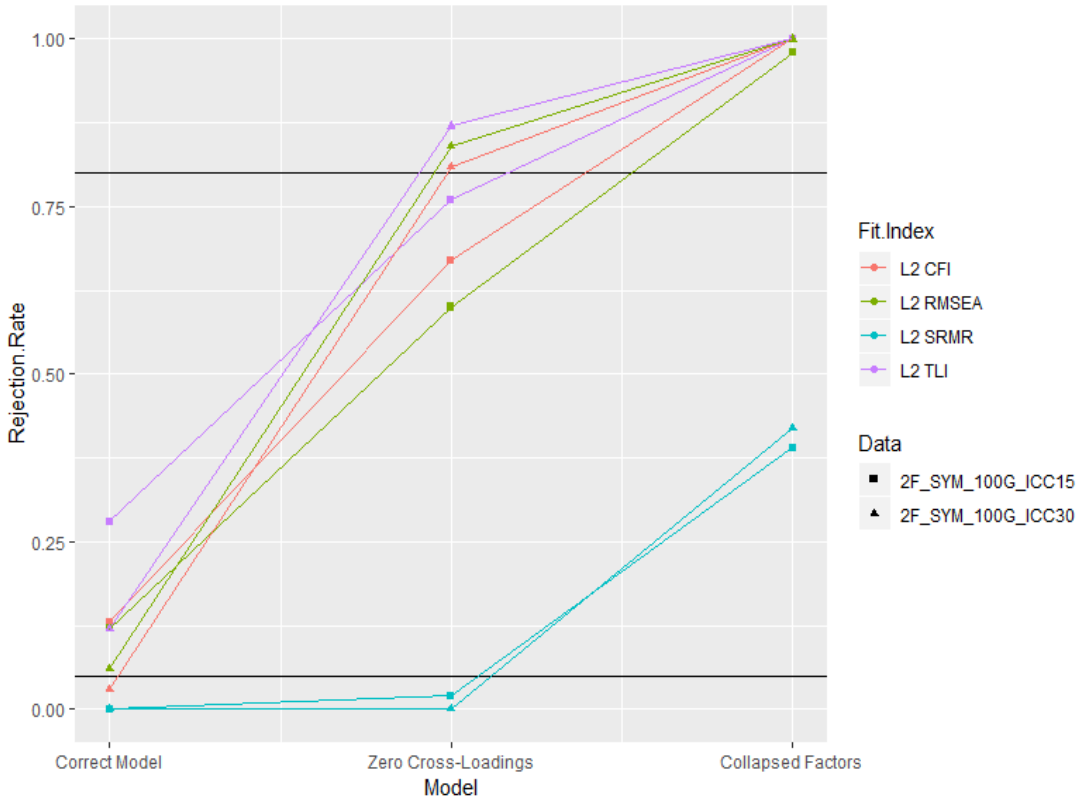

Figure 37. Impact of ICCs on Level-2 Fit Indices' Rejection Rates with Small Models, Symmetric Data, and Many Groups.

Note. This figure has overlap. For correct models, Level-2 SRMR rejection rates were zero with small and large ICCs. For correct models, rejection rates were identical for Level-2 RMSEA with small ICCs and Level-2 TLI with large ICCs; they differed .01 from Level-2 CFI's rejection rate with small ICCs. For collapsed factors, rejection rates were 1 for Level-2 CFI and TLI with small and large ICCs and for Level-2 RMSEA with large ICCs.

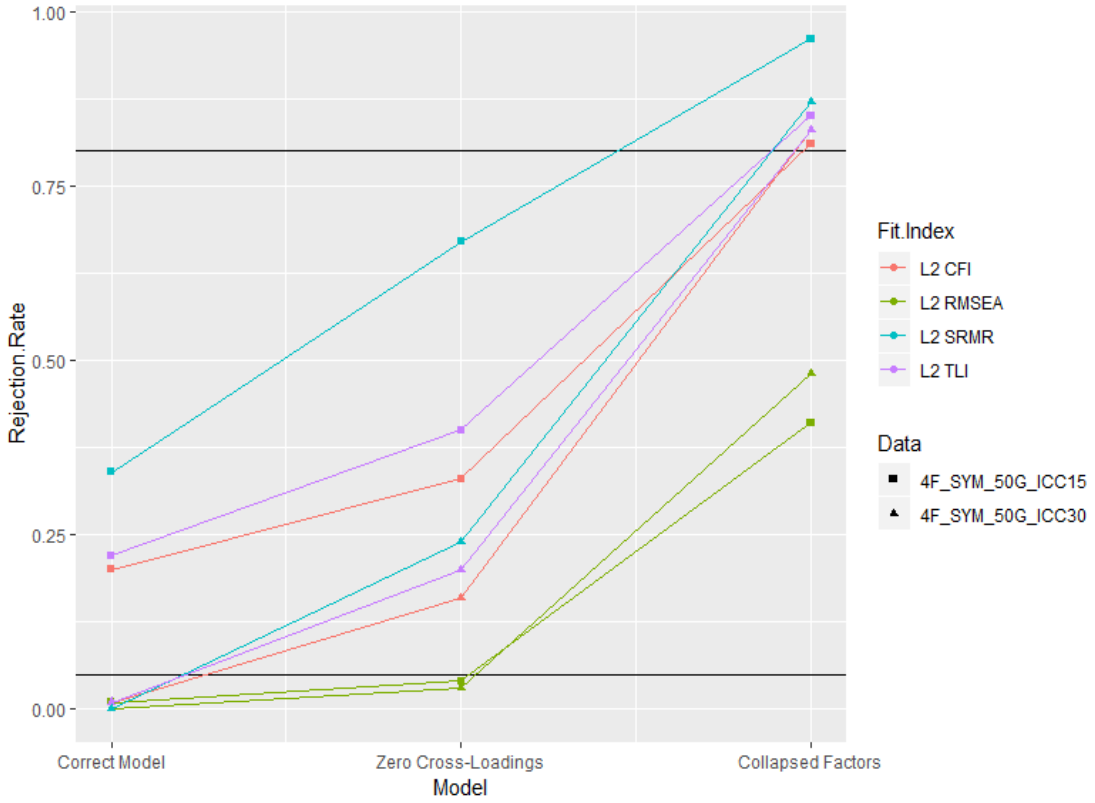

Figure 38. Impact of ICCs on Level-2 Fit Indices' Rejection Rates with Large Models, Symmetric Data, and Few Groups.
Note. This figure has overlap. For correct models, Level-2 RMSEA and SRMR's rejection rates were zero with large ICCs. For correct models, rejection rates were .01 for Level-2 RMSEA with small ICCs and for Level-2 CFI and TLI with large ICCs. For collapsed factors, Level-2 CFI and TLI's rejection rates were identical with large ICCs.

Level-2 fit indices' rejection rates for Level-2 cross-loadings fixed to 0. ICCs differentially impacted Level-2 CFI rejection of Level-2 cross-loadings fixed to 0 based on model size. For large models, Level-2 CFI rejection rates usually *decreased* as ICCs increased. These situations yielded rejection rates of .33 to .49 with small ICCs and .16 to .47 with large ICCs. With large models, Level-2 CFI performed worst with few groups, yielding rejection rates of .33 with small ICCs and .16 to .20 with large ICCs. With large models, Level-2 CFI performed best with symmetric data and many groups. These situations produced rejection rates of .49 with few groups and .47 with many groups. For large models, asymmetric data, and many groups, rejection rates increased from .40 with small ICCs to .43 with large ICCs.

For small models, Level-2 CFI rejection of Level-2 cross-loadings fixed to 0 usually *increased* as ICCs increased. These situations yielded rejection rates of .45 to .67 with small ICCs and .50 to .81 with large ICCs. For small models, Level-2 CFI performed best with symmetric data and many groups. These situations yielded rejection rates of .67 with small ICCs and .81 with large ICCs. For small models, Level-2 CFI performed worst with few groups. These situations produced rejection rates of .45 to .52 with small ICCs and .46 to .50 with large ICCs. For small models, symmetric data, and few groups, rejection rates decreased from .52 with small ICCs to .46 with large ICCs.

For large models, Level-2 TLI rejection rates usually *decreased* as ICCs increased. These situations yielded rejection rates of .35 to .57 with small ICCs and .20 to .54 with large ICCs. For large models, ICCs most impacted Level-2 TLI with asymmetric data, few groups, and small ICCs. These situations yielded rejection rates of .40 with

small ICCs and .20 with large ICCs. For large models, Level-2 TLI performed worst with few groups, yielding rejection rates of .35 to .40 with small ICCs and .20 to .24 with large ICCs. With large models, Level-2 TLI performed best with symmetric data and many groups. These situations produced rejection rates of .57 with few groups and .54 with many groups. For large models, asymmetric data, and many groups, Level-2 TLI rejection rates increased from .43 with small ICCs to .54 with large ICCs.

For small models, Level-2 TLI rejection of Level-2 cross-loadings fixed to 0 usually *increased* as ICCs increased. These situations yielded rejection rates of .52 to .76 with small ICCs and .59 to .87 with large ICCs. For small models, Level-2 TLI performed best with symmetric data and many groups. These situations yielded rejection rates of .76 with small ICCs and .87 with large ICCs. For small models, Level-2 TLI performed worst with few groups. These situations produced rejection rates of .52 to .56 with small ICCs and .55 to .59 with large ICCs. For small models, symmetric data, and few groups, rejection rates decreased from .56 with small ICCs to .55 with large ICCs.

Level-2 RMSEA rejection of Level-2 cross-loadings fixed to 0 usually increased as ICCs increased. Rejection rates were .02 to .60 with small ICCs and .06 to .84 with large ICCs. Level-2 RMSEA performed worst with large models, asymmetric data, and many groups. These situations produced rejection rates of .02 with small ICCs and .06 with large ICCs. Level-2 RMSEA performed best with small models, symmetric data, and many groups. These situations yielded rejection rates of .60 with small ICCs and .84 with large ICCs. With large models, asymmetric data, and few groups, Level-2 RMSEA rejection rates were .05 regardless of ICCs. Level-2 RMSEA rejection rates decreased as

ICCs increased in two situations. For small models, symmetric data, and few groups, Level-2 RMSEA rejection rates were .49 with small ICCs and .48 with large ICCs. For large models, symmetric data, and few groups, Level-2 RMSEA rejection rates were .04 with small ICCs and .03 with large ICCs.

Level-2 SRMR rejection of Level-2 cross-loadings fixed to 0 almost always decreased as ICCs increased. Rejection rates were .01 to .89 with small ICCs and 0 to .41 with large ICCs. Level-2 SRMR performed best with large models, severe asymmetry, and few groups. These situations yielded rejection rates of .89 with small ICCs and .41 with large ICCs. Level-2 SRMR performed worst with small models and many groups. These situations produced rejection rates of .01 to .02 with small ICCs and 0 with large ICCs. With large models, symmetric data, and many groups, Level-2 SRMR rejection rates were 0 regardless of ICCs.

Level-2 fit indices' rejection rates for Level-2 collapsed factors. Level-2 CFI rejection of Level-2 collapsed factors almost always increased as ICCs increased. Rejection rates were .75 to .98 with small ICCs and .83 to 1 with large ICCs. Level-2 CFI performed worst with large models and few groups. These situations produced rejection rates of .75 to .81 with small ICCs and .83 to .84 with large ICCs. Level-2 CFI performed best with large models, symmetric data, and many groups. These situations yielded rejection rates of .98 with small ICCs and 1 with large ICCs. Level-2 CFI rejection rates were 1 regardless of ICCs for small models, symmetric data, and many groups.

Level-2 TLI rejection of Level-2 collapsed factors usually increased as ICCs increased. Rejection rates were .77 to .99 with small ICCs and .85 to 1 with large ICCs.

Level-2 TLI performed worst with large models, asymmetric data, and few groups. These situations produced rejection rates of .77 with small ICCs and .85 with large ICCs. Level-2 TLI performed best with large models, symmetric data, and many groups. These situations yielded rejection rates of .99 with small ICCs and 1 with large ICCs. Level-2 TLI rejection rates were 1 regardless of ICCs with small models, symmetric data, and many groups. For large models, symmetric data, and few groups, rejection rates decreased from .85 with small ICCs to .83 with large ICCs.

Level-2 RMSEA rejection of Level-2 collapsed factors always increased as ICCs increased. Rejection rates were .24 to .98 with small ICCs and .48 to 1 with large ICCs. Level-2 RMSEA performed best with small models, symmetric data, and many groups, yielding rejection rates of .98 with small ICCs and 1 with large ICCs. With small ICCs, Level-2 RMSEA performed worst with large models, asymmetry, and few groups (rejection rate = .24). With large ICCs, Level-2 RMSEA performed worst with large models, symmetry, and few groups (rejection rate = .48). ICCs most affected Level-2 RMSEA with large models, severe asymmetry, and many groups. This situation produced rejection rates of .45 with small ICCs and .89 with large ICCs (increase of .44). ICCs least affected Level-2 RMSEA with small models, symmetric data, and many groups. This situation yielded rejection rates of .98 with small ICCs and 1 with large ICCs.

Level-2 SRMR rejection rates of Level-2 collapsed factors usually *decreased* as ICCs increased. Level-2 SRMR rejection rates were .39 to .99 with small ICCs and .41 to .95 with large ICCs. Level-2 SRMR performed best with large models and few groups. These conditions produced rejection rates of .96 to .99 with small ICCs and .87 to .95

with large ICCs. Level-2 SRMR performed worst with small models and many groups, yielding rejection rates of .39 to .41 with small ICCs and .41 to .42 with large ICCs.

Level-2 fit indices' rejection rates of Level-2 correct models. Level-2 CFI rejection rates of Level-2 correct models always decreased as ICCs increased. Level-2 CFI rejection rates were .08 to 22 with small ICCs and .01 to .16 with large ICCs. For large ICCs, Level-2 CFI performed best with large models, symmetric data, and few groups. These conditions yielded correct model rejection rates of .01. Level-2 CFI performed worst with small models, severe asymmetry, and few groups. These conditions produced correct model rejection rates of .22 with small ICCs and .16 with large ICCs. Otherwise, with large ICCs, Level-2 CFI rejection rates were .01 to .07 for correct models. Level-2 CFI maintained correct model rejection rates (< .05) only with many groups and large ICCs. Level-2 CFI produced correct model rejection rates of .01 with large models, symmetry, few groups, and large ICCs. ICCs most impacted Level-2 CFI with large models, symmetric data, and few groups. This situation yielded rejection rates of .20 with small ICCs and .01 with large ICCs.

Level-2 TLI rejection rates of Level-2 correct models always decreased as ICCs increased. Level-2 TLI rejection rates were .15 to 28 with small ICCs and .01 to .18 with large ICCs. Level-2 TLI performed worst with small models. These conditions produced correct model rejection rates of .26 to .28 with small ICCs and .12 to .18 with large ICCs. With small ICCs, Level-2 TLI performed best with large models, symmetry, and many groups (rejection rate = .13). With large ICCs, Level-2 TLI performed best with large models, symmetry, and few groups (rejection rate = .01). With large ICCs and large

models, Level-2 TLI rejection rates were .01 to .07 for correct models. ICCs most impacted Level-2 TLI with large models, symmetry, and few groups. These situations yielded rejection rates of .22 with small ICCs and .01 with large ICCs.

Level-2 RMSEA rejection rates of Level-2 correct models usually decreased as ICCs increased. Level-2 RMSEA rejection rates were .01 to .20 with small ICCs and 0 to .15 with large ICCs. Level-2 RMSEA rejected small correct models much more frequently than large correct models. For small correct models, Level-2 RMSEA rejection rates were .10 to .20 with small ICCs and .06 to .15 with large ICCs. Level-2 RMSEA rejection rates of large correct models were 0 to .01 with small ICCs and always 0 with large ICCs.

Level-2 SRMR rejection rates of Level-2 correct models either decreased as ICCs increased or remained zero regardless of ICCs. Level-2 SRMR rejection rates were .01 to .80 with small ICCs and 0 to .08 with large ICCs. With small ICCs, Level-2 SRMR performed worst with large models and few groups. These situations produced rejection rates of .34 with symmetric data and .80 with asymmetric data. Otherwise, with small ICCs, Level-2 SRMR rejection rates were 0 to .09. With large ICCs, Level-2 SRMR rejected correct models most frequently (rejection = .08) with large models, severe asymmetry, and few groups. Otherwise, with large ICCs, Level-2 SRMR never rejected Level-2 correct models.

Aggregate Fit Indices' Overall Performance

Table 8 provides aggregate fit index rejection rates for all small model conditions. Table 9 provides aggregate fit index rejection rates for all large model conditions

Generally, aggregate fit indices identified Level-1 misfit much more accurately than Level-2 misfit. Aggregate CFI and TLI rejected 90-100% of models with collapsed Level-1 factors. Conversely, aggregate CFI, TLI, and RMSEA rejected 0-1% of models with only Level-2 misfit. Aggregate RMSEA rejection of Level-1 collapsed models varied widely for small models but never rejected large models. Generally, aggregate fit indices rejection rates for small models were slightly lower for misfit at both levels compared to misfit at Level-1 only. Aggregate fit indices' rejection rates for large models usually stayed the same for misfit at both levels compared to misfit at Level-1 only. Aggregate fit indices never rejected the correct model regardless of data condition.

Aggregate CFI and TLI performed much better than aggregate RMSEA. Aggregate RMSEA never rejected large models for any misspecification condition. Aggregate CFI and TLI generally performed similarly. Aggregate TLI usually had slightly higher rejection rates than CFI. TLI did much better than CFI for small models with Level-1 cross-loadings fixed to 0. In these situations, aggregate TLI rejection rates ranged from .33 to .85, whereas CFI rejection rates ranged from .10 to .18. Aggregate TLI performed best with small models, symmetric data, many groups, and small ICCs (rejection rate = .85). For small models, aggregate TLI performed worst with asymmetric data and few groups (rejection rates = .33 to .43). Aggregate fit indices rejection rates generally were lower with large ICCs than small ICCs.

Level-1 cross-loadings fixed to 0. Aggregate fit indices' rejection rates for Level-1 cross-loadings fixed to 0 were always poor for large models and usually poor for small models. Aggregate CFI rejection rates were 0 to .03 for large models and .10 to .18

for small models. TLI rejection rates were .01 to .04 for large models and .33 to .85 for small models. TLI performed best with small models, symmetric data and many groups (rejection rates = .78 to .85). For small models, aggregate TLI performed worst with asymmetric data and few groups (rejection rates = .33 to .43). Aggregate RMSEA rejection rates were 0 for large models and 0 to .10 for small models.

Level-1 collapsed factors. Aggregate CFI and TLI rejected 90% to 100% of Level-1 collapsed factors, whereas RMSEA rejection rates depended on model size. Aggregate RMSEA rejection rates were always zero for large models and .10 to 1 for small models. For small models, aggregate RMSEA performed worst with asymmetric data and large ICCs (rejection rates = .10 to .27). For small models, aggregate RMSEA performed best with symmetric data (rejection rates = .92 to 1).

Level-2 cross-loadings fixed to 0. Aggregate CFI, TLI, and RMSEA never rejected Level-2 cross-loadings fixed to 0.

Level-2 collapsed factors. Aggregate CFI, TLI, and RMSEA rejected 0 to 1% of Level-2 collapsed factors.

Level-1 and Level-2 cross-Loadings fixed to 0. Rejection of cross-loadings fixed to 0 at both levels was always poor for CFI and RMSEA but sometimes good for TLI with small models. Aggregate CFI rejection rates were .01 to .02 for small models and .07 to .21 for large models. Aggregate RMSEA rejection rates were 0 for all large models and 0 to .08 for small models. Aggregate TLI rejection rates were .01 to .04 for large models and .34 to .84 for small models. Aggregate TLI performed best with

symmetric data and many groups (rejection rates = .83 to .84). Aggregate TLI performed worst with few groups and large ICCs (rejection rates = .34 to .37).

Level-1 and Level-2 collapsed factors. Aggregate CFI and TLI rejection rates for collapsed factors at both levels were excellent, whereas RMSEA performance varied widely and depended on model size. Aggregate CFI rejected 100% of small models and 94% to 100% of large models. Aggregate TLI always rejected 100% of models regardless of model size. Aggregate RMSEA rejected 0% of large models and 9% to 100% of small models. For small models, aggregate RMSEA performed worst with asymmetric data and large ICCs (rejection rates = .09 to .24). For small models, aggregate RMSEA performed best with symmetric data (rejection rates = .90 to 1). Notably, aggregate RMSEA rejected 91% of small models fit to asymmetric data with many groups and small ICCs.

Correct model at both levels. Aggregate CFI, TLI, and RMSEA never rejected the correct model regardless of data condition.

Table 8. Rejection Rates of Aggregate Fit Indices for Small Model Conditions

Fit Index	L1 Misp. Cross-loadings	L1 Misp. Factors	L2 Misp. Cross-loadings	L2 Misp. Factors	L1 & L2 Misp. Cross-loadings	L1 & L2 Misp. Factors	True Model	Data
CFI	0.17	1	0	0	0.17	1	0	Skewed data, 50 groups, ICC = .15
TLI	0.43	1	0	0	0.41	1	0	
RMSEA	0	0.73	0	0	0	0.65	0	
CFI	0.1	1	0	0	0.1	1	0	Skewed data, 100 groups, ICC = .15
TLI	0.64	1	0	0	0.63	1	0	
RMSEA	0	0.92	0	0	0	0.91	0	
CFI	0.13	1	0	0	0.15	1	0	Skewed data, 50 groups, ICC = .30
TLI	0.33	1	0	0	0.34	1	0	
RMSEA	0	0.1	0	0	0	0.09	0	
CFI	0.12	1	0	0	0.16	1	0	Skewed data, 100 groups, ICC = .30
TLI	0.59	1	0	0.01	0.65	1	0	
RMSEA	0	0.27	0	0	0	0.24	0	
CFI	0.18	1	0	0	0.18	1	0	Normal data, 50 groups, ICC = .15
TLI	0.63	1	0	0	0.59	1	0	
RMSEA	0.08	1	0	0	0.06	1	0	
CFI	0.17	1	0	0	0.2	1	0	Normal data, 100 groups, ICC = .15
TLI	0.85	1	0	0	0.84	1	0	
RMSEA	0.1	1	0	0	0.08	1	0	
CFI	0.07	1	0	0	0.07	1	0	Normal data, 50 groups, ICC = .30
TLI	0.36	1	0	0	0.37	1	0	
RMSEA	0	0.92	0	0	0	0.9	0	
CFI	0.16	1	0	0	0.21	1	0	Normal data, 100 groups, ICC = .30
TLI	0.78	1	0	0	0.83	1	0	
RMSEA	0.01	1	0	0	0.01	1	0	

Note. Misp = misspecified.

Table 9. Rejection Rates of Aggregate Fit Indices for Large Model Conditions

Fit Index	L1 Misp. Cross-loadings	L1 Misp. Factors	L2 Misp. Cross-loadings	L2 Misp Factors	L1 & L2 Misp Cross-loadings	L1 & L2 Misp Factors	True Model	Data
CFI	0.02	0.97	0	0	0.02	0.98	0	Skewed data, 50 groups, ICC = .15
TLI	0.03	1	0	0	0.02	1	0	
RMSEA	0	0	0	0	0	0	0	
CFI	0	0.99	0	0	0	0.99	0	Skewed data, 100 groups, ICC = .15
TLI	0.03	1	0	0	0.03	1	0	
RMSEA	0	0	0	0	0	0	0	
CFI	0.02	0.9	0	0	0.01	0.94	0	Skewed data, 50 groups, ICC = .30
TLI	0.04	1	0	0	0.04	1	0	
RMSEA	0	0	0	0	0	0	0	
CFI	0.01	1	0	0	0.01	1	0	Skewed data, 100 groups, ICC = .30
TLI	0.02	1	0	0	0.04	1	0	
RMSEA	0	0	0	0	0	0	0	
CFI	0.01	1	0	0	0.01	1	0	Normal data, 50 groups, ICC = .15
TLI	0.01	1	0	0	0.01	1	0	
RMSEA	0	0	0	0	0	0	0	
CFI	0.01	1	0	0	0.01	1	0	Normal data, 100 groups, ICC = .15
TLI	0.05	1	0	0	0.05	1	0	
RMSEA	0	0	0	0	0	0	0	
CFI	0.01	0.95	0	0	0.01	0.97	0	Normal data, 50 groups, ICC = .30
TLI	0.01	1	0	0	0.01	1	0	
RMSEA	0	0	0	0	0	0	0	
CFI	0.01	1	0	0	0.01	1	0	Normal data, 100 groups, ICC = .30
TLI	0.04	1	0	0	0.04	1	0	
RMSEA	0	0	0	0	0	0	0	

Note. Misp = misspecified.

Data Asymmetry's Impact on Aggregate Fit Indices

Before describing the impact of data asymmetry on aggregate fit indices' rejection rates, some figures will be shown. These figures introduce the complex nature of aggregate fit index performance. Data asymmetry's impact generally also depended on the specific fit index, type of misspecification, level(s) modeled, ICCs, model size, and number of groups. Figures 39 and 40 contrast aggregate fit indices' rejection rates with symmetric and severely asymmetric data for various combinations of conditions. Tables 8 and 9 provide aggregate fit indices' rejection rates.

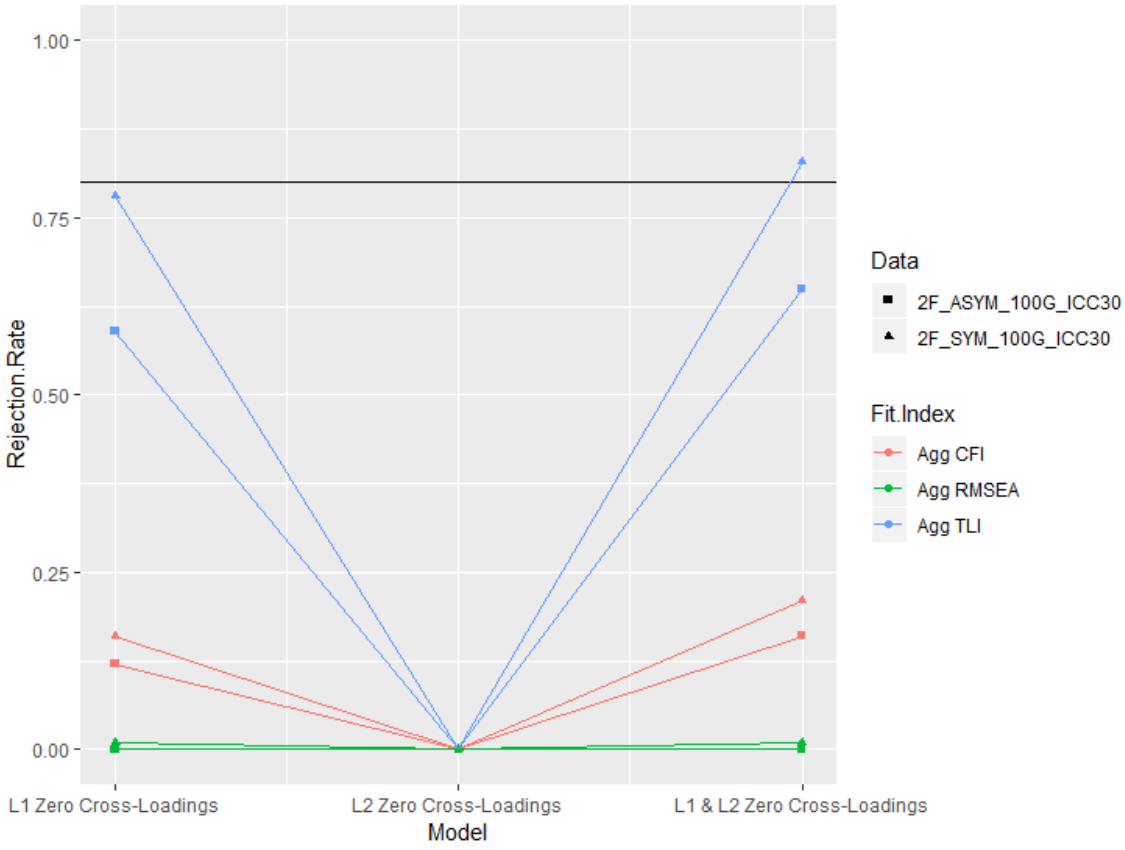

Figure 39. Impact of Data Asymmetry on Aggregate Fit Indices' Rejection of Cross-Loadings Fixed to Zero with Small Models, Many Groups, and Large ICCs.
Note. This figure has overlap. All aggregate fit indices' rejection rates were zero for Level-2 cross-loadings fixed to zero regardless of asymmetry.

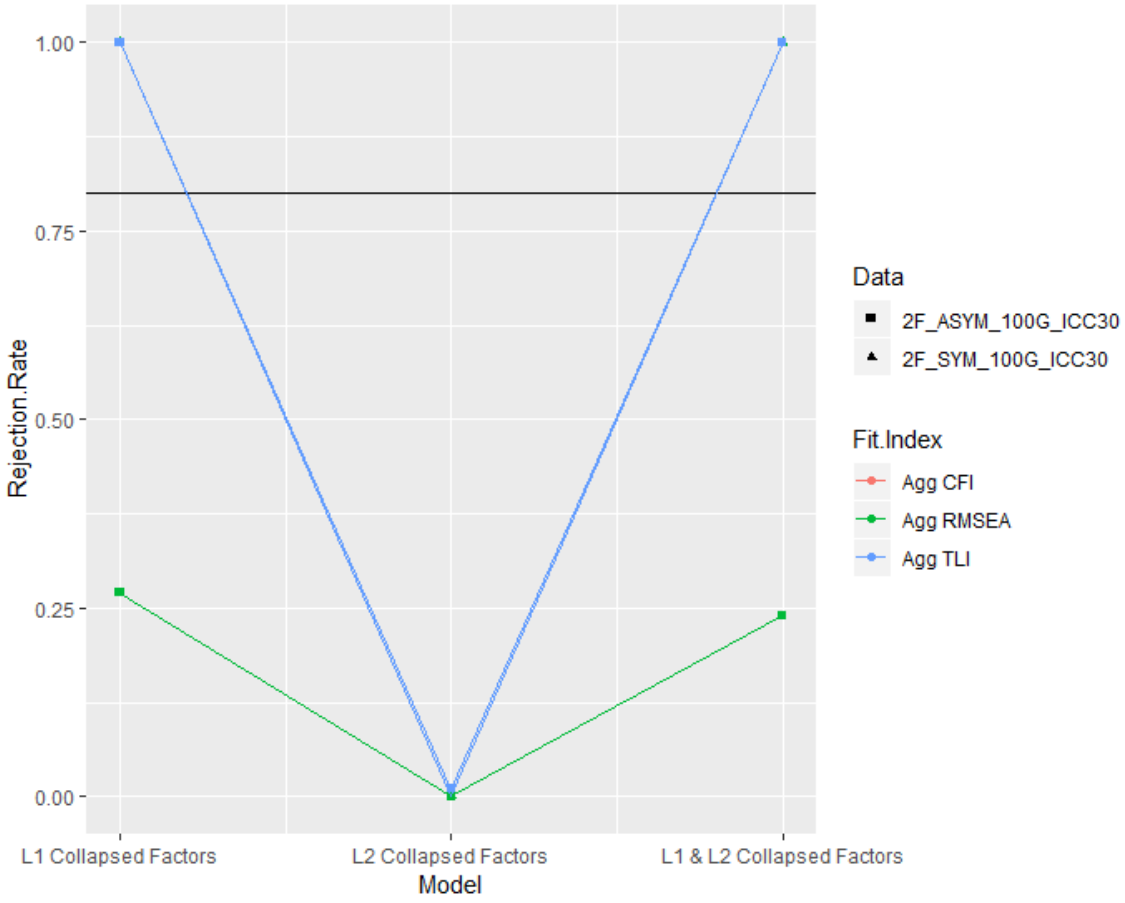

Figure 40. Impact of Data Asymmetry on Aggregate Fit Indices' Rejection of Collapsed Factors with Small Models, Many Groups, and Large ICCs.
Note. Overlap occurred in this figure. Agg CFI lines are occluded by the Agg TLI lines; Agg CFI values are identical to Agg TLI values.

Level-1 cross-loadings fixed to 0. Data asymmetry differentially affected aggregate fit indices' rejection of Level-1 cross-loadings fixed to 0 based on model size. For small models, aggregate CFI rejection rates usually decreased as asymmetry increased. Small model rejection rates were .16 to .18 with symmetric data and .12 to .17 with asymmetric data. However, CFI rejection rates were .07 with symmetry and .13 with asymmetry given small models, symmetric data, few groups, and large ICCs. For large models, aggregate CFI rejection rates changed by 0 to .01 as asymmetry increased. For large models, all aggregate CFI rejection rates were .01 with symmetric data and 0 to .02 with asymmetric data.

For small models, aggregate TLI rejection rates always decreased as asymmetry increased. Rejection rates were .36 to .85 with symmetry and .33 to .64 with asymmetry. For small models, aggregate TLI performed worst with few groups and large ICCs, yielding rejection rates of .36 with symmetric data and .33 with asymmetric data. For small models, aggregate TLI performed best with many groups and small ICCs. In these situations, rejection rates were .85 with symmetric data and .64 with asymmetric data.

For large models, asymmetry affected aggregate TLI rejection rates differently based on number of groups. With few groups, TLI rejection rates increased as asymmetry increased, yielding rejection rates of .01 with symmetry and .03 to .04 with asymmetry. With many groups, TLI rejection rates decreased as asymmetry increased, yielding rejection rates of .04 to .05 with symmetry and .02 to .03 with asymmetry.

Data asymmetry affected aggregate RMSEA rejection of Level-1 cross-loadings fixed to 0 differently based on model size. Aggregate RMSEA never rejected large

models regardless of asymmetry. For small models, RMSEA rejection usually decreased as asymmetry increased. Rejection rates were .08 to .10 with symmetry and 0 with asymmetry.

Collapsed Level-1 factors. Data asymmetry weakly affected aggregate CFI rejection of collapsed Level-1 factors with large models and did not affect CFI with small models. With small models, aggregate CFI rejected 100% of collapsed Level-1 factors regardless of asymmetry. With large models, CFI rejection usually decreased as asymmetry increased. Rejection rates were .95 to 1 with symmetry and .9 to 1 with asymmetry.

Aggregate TLI always rejected 100% of collapsed Level-1 factors regardless of asymmetry or model size.

Data asymmetry differentially affected aggregate RMSEA rejection of collapsed Level-1 factors based on model size. RMSEA never rejected large collapsed models regardless of asymmetry. With small models, RMSEA rejection rates always decreased as asymmetry increased. Rejection rates were .92 to 1 with symmetry and .1 to .92 with asymmetry. Asymmetry most strongly affected RMSEA rejection rates with large ICCs, yielding rejection rates of .92 to 1 with symmetry and .10 to .27 with asymmetry. For small models and small ICCs, aggregate RMSEA rejection rates were less affected by asymmetry, yielding rejection rates of 1 with symmetry and .73 to .92 with asymmetry.

Level-2 cross-loadings fixed to 0. Aggregate CFI, TLI, and RMSEA never rejected Level-2 cross-loadings fixed to 0 regardless of asymmetry.

Collapsed Level-2 factors. Aggregate CFI, TLI, and RMSEA almost never rejected collapsed Level-2 factors regardless of asymmetry. However, as asymmetry increased, aggregate TLI rejection increased from 0 to .01 with small models, severe asymmetry, many groups, and large ICCs. Otherwise, aggregate fit indices never rejected collapsed Level-2 factors regardless of asymmetry.

Cross-loadings fixed to 0 at both levels. Asymmetry affected aggregate CFI, TLI, and RMSEA rejection of cross-loadings fixed to 0 at both levels different based on model size.

For small models, aggregate CFI rejection of cross-loadings fixed to 0 at both levels usually decreased as asymmetry increased. Rejection rates were .16 to .18 with symmetry and .12 to .17 with asymmetry. However, CFI rejection rates increased as asymmetry increased with few groups and large ICCs. In this situation, CFI rejection rates increased from .07 with symmetry to .13 with asymmetry.

For large models, aggregate CFI rejection of cross-loadings fixed to 0 at both levels was essentially zero regardless of asymmetry. CFI rejection rates were .01 with symmetry and 0 to .02 with asymmetry.

For small models, aggregate TLI rejection of cross-loadings fixed to 0 at both levels always decreased as asymmetry increased. Rejection rates were .37 to .84 with symmetry and .34 to .65 with asymmetry. For small models, data asymmetry most strongly affected TLI with many groups and small ICCs. In this situation, TLI rejection rates were .85 with symmetry and .64 with asymmetry. Data asymmetry least affected

TLI with few groups and large ICCs. In this situation, TLI rejection rates were .36 with symmetry and .33 with asymmetry.

With large models, asymmetry had a weak, inconsistent effect on aggregate TLI rejection of cross-loadings fixed to 0 at both levels. Large model rejection rates were .01 to .05 with symmetry and .02 to .04 with asymmetry. With few groups, TLI rejection rates increased as asymmetry increased.

Aggregate RMSEA never rejected large models with cross-loadings fixed to 0 at both levels regardless of asymmetry. For small models, aggregate RMSEA rejection rates usually decreased as asymmetry increased. Rejection rates were .01 to .08 with symmetry and 0 with asymmetry. With small models, few groups, and large ICCs, aggregate RMSEA rejection rates were 0 regardless of asymmetry.

Collapsed factors at both levels. Asymmetry weakly affected aggregate CFI rejection of collapsed factors at both levels. CFI rejection rates were .97 to 1 with symmetry and .94 to 1 with asymmetry. Asymmetry slightly affected CFI rejection rates with large models. With large models, CFI rejection rates decreased as asymmetry increased. Rejection rates were .97 to 1 with symmetry and .94 to 1 with asymmetry. For large models, asymmetry most affected CFI rejection rates with few groups and large ICCs. In this situation, rejection rates were .97 with symmetry and .94 with asymmetry. Otherwise, CFI rejection rates were 1 with symmetry and .98 to 1 with asymmetry.

Aggregate TLI always rejected 100% of collapsed factors at both levels regardless of asymmetry.

Data asymmetry affected aggregate RMSEA rejection of collapsed factors at both levels differently based on model size. Aggregate RMSEA never rejected large collapsed models regardless of asymmetry. For small collapsed models, RMSEA rejection rates usually decreased as asymmetry increased. Rejection rates were .90 to 1 with symmetry and .09 to .91 with asymmetry. Asymmetry least affected RMSEA with many groups and small ICCs. In this situation, rejection rates were 1 with symmetry and .91 with asymmetry. Asymmetry most affected aggregate RMSEA with few groups and large ICCs. In this situation, rejection rates were .9 with symmetry and .09 with asymmetry.

Correct model at both levels. Aggregate CFI, TLI, and RMSEA never rejected the correct model at both levels regardless of asymmetry.

Model Size's Impact on Aggregate Fit Indices

Before describing the impact of model size on aggregate fit indices' rejection rates, some figures will be shown. These figures introduce the complex nature of aggregate fit index performance. Model size's impact generally also depended on the specific fit index, type of misspecification, level(s) modeled, asymmetry, ICCs, and number of groups. Figures 41 and 42 contrast aggregate fit indices' rejection rates with small and large models for various combinations of conditions. Tables 8 and 9 give aggregate fit indices' rejection rates.

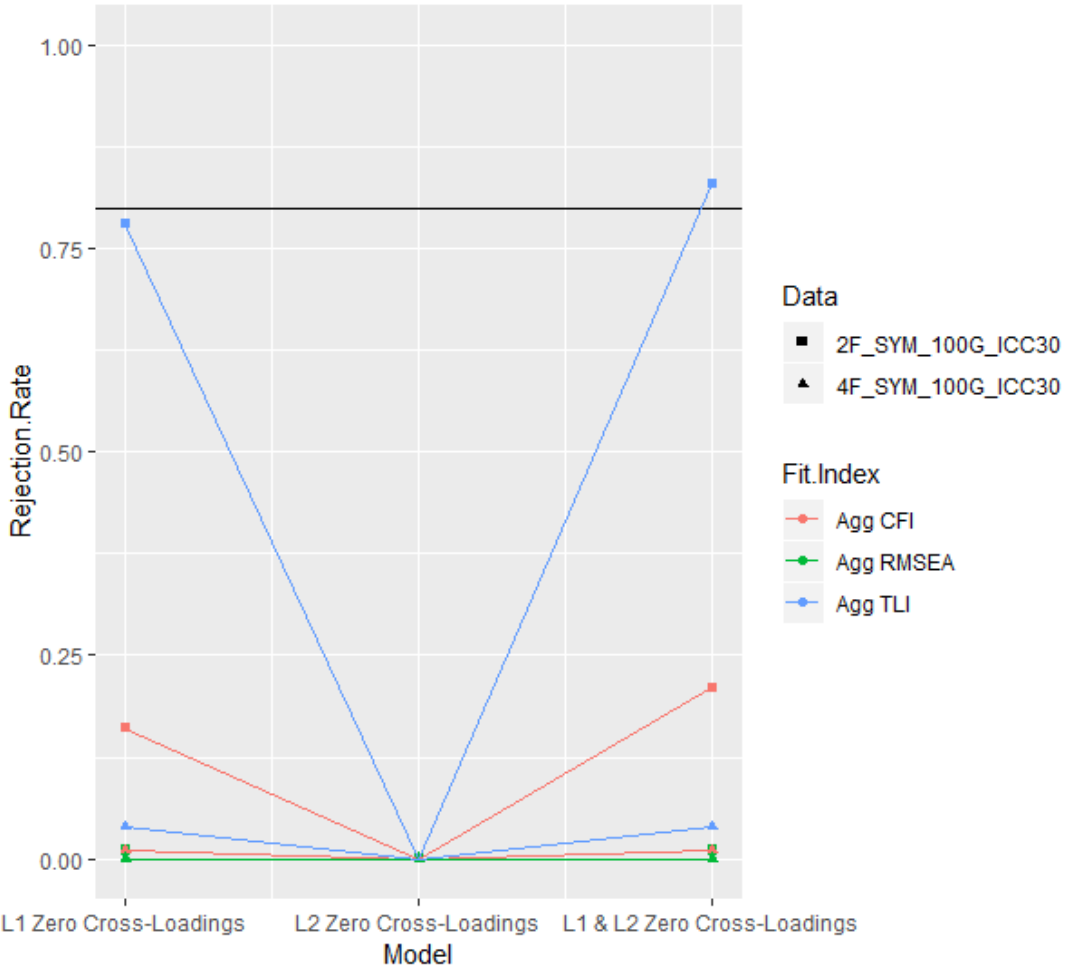

Figure 41. Impact of Model Size on Aggregate Fit Indices' Rejection of Cross-Loadings with Symmetric Data, Many Groups, and Large ICCs.

Note. This figure has overlap. For Level-1 cross-loadings, rejection rates were .01 for aggregate RMSEA with small models and aggregate CFI with large models. For Level-2 cross-loadings, rejection rates were always zero for all aggregate fit indices. For zero Level-1 and Level-2 cross-loadings, rejection rates were .01 for aggregate RMSEA with small models and aggregate CFI with large models.

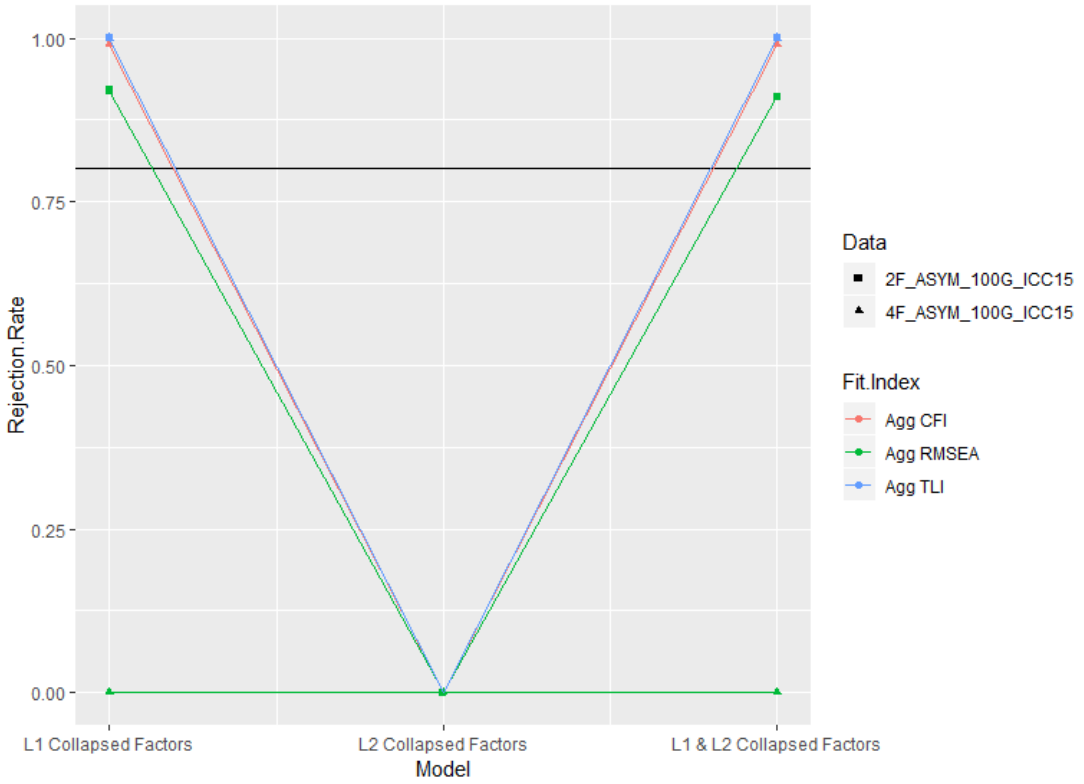

Figure 42. Impact of Model Size on Aggregate Fit Indices' Rejection of Collapsed Factors with Asymmetric Data, Many Groups, and Small ICCs.
Note. This figure has overlap. All aggregate fit indices' rejection rates were zero for Level-2 collapsed factors. Aggregate TLI and CFI always overlapped each other, as they always had identical or nearly identical (.01 difference) rejection rates.

Level-1 cross-loadings fixed to 0. Aggregate CFI rejection of Level-1 cross-loadings fixed to 0 always decreased as model size increased. Rejection rates were .07 to .18 for small models and 0 to .02 for large models.

Aggregate TLI rejection rates for Level-1 cross-loadings fixed to 0 always decreased as model size increased. Rejection rates were .36 to .85 with small models and .01 to .05 with large models. TLI performed best with small models, many groups, and symmetric data, producing rejection rates of .78 to .85.

With asymmetric data, aggregate RMSEA never rejected Level-1 cross-loadings fixed to 0 regardless of model size. With symmetric data, RMSEA rejection rates usually decreased as model size increased. Rejection rates were .01 to .10 with small models and 0 with large models. RMSEA never rejected models with few groups and large ICCs regardless of model size.

Collapsed Level-1 factors. Aggregate CFI rejection of collapsed Level-1 factors sometimes decreased as model size increased. These decreases usually occurred with asymmetric data or symmetric data with few groups and large ICCs. Rejection rates were 1 with small models and .90 to .99 with large models. Model size most affected CFI rejection rates with few groups and large ICCs, yielding rejection rates of .90 to .95 with large models. Otherwise, rejection rates were .97 to 1 with large models.

Aggregate TLI always rejected 100% of collapsed Level-1 factors regardless of model size.

Aggregate RMSEA rejection of collapsed Level-1 factors always decreased as model size increased. Rejection rates were .73 to 1 with small models and 0 with large

models. Model size most impacted RMSEA with symmetric data, as rejection rates were .92 to 1 with small models and 0 with large models.

Level-2 cross-loadings fixed to 0. Aggregate CFI, TLI, and RMSEA never rejected Level-2 cross-loadings fixed to 0 regardless of model size.

Collapsed Level-2 factors. Aggregate CFI, TLI, and RMSEA almost never rejected collapsed Level-2 factors regardless of model size. However, with asymmetry, many groups, and large ICCs, TLI rejected 1% of collapsed small models and 0% of collapsed large models.

Cross-loadings fixed to 0 at both levels. Aggregate CFI rejection of cross-loadings fixed to 0 at both levels always decreased as model size increased. Rejection rates were .07 to .21 with small models and 0 to .02 with large models.

Aggregate TLI rejection of cross-loadings fixed to 0 at both levels always decreased as model size increased. Rejection rates were .34 to .84 with small models and .01 to .05 with large models. Model size most affected TLI with many groups and symmetric data. In these situations, rejection rates were .83 to .84 with small models and .04 to .05 with large models.

Model size differentially affected aggregate RMSEA rejection of cross-loadings fixed to 0 at both levels based on data asymmetry. With asymmetric data, aggregate RMSEA rejection rates were always 0 regardless of model size. With symmetric data, RMSEA rejection rates usually decreased as model size increased. Rejection rates were .01 to .08 with small models and 0 with large models.

Aggregate fit indices' rejection of collapsed factors at both levels. Model size differentially affected aggregate CFI rejection of collapsed factors at both levels based on data asymmetry. With asymmetric data, CFI rejection rates usually decreased as model size increased. Rejection rates were 1 with small models and .94 to 1 with large models. With symmetric data, aggregate CFI usually rejected 100% of collapsed models regardless of model size. However, with few groups and large ICCs, rejection rates were 1 with small models and .97 with large models.

Aggregate TLI always rejected 100% of collapsed factors at both levels regardless of model size.

Aggregate RMSEA rejection of collapsed factors at both levels always decreased as model size increased. Rejection rates were .09 to 1 with small models and 0 with large models. With small models, RMSEA performed worst with severe asymmetry, few groups, and large ICCs. In these situations, rejection rates were .09 to .24. With small models, RMSEA performed best with symmetry, many groups, and small ICCs. These situations yielded rejection rates of 1.

Correct model at both levels. Aggregate CFI, TLI, and RMSEA never rejected the correct model at both levels regardless of model size.

Impact of Number of Groups on Aggregate Fit Indices

Before describing the impact of number of groups on aggregate fit indices' rejection rates, some figures will be shown. These figures introduce the complex nature of aggregate fit index performance. The impact of number of groups generally also depended on the specific fit index, type of misspecification, level(s) modeled, ICCs,

model size, and data asymmetry. Figures 43 and 44 contrast aggregate fit indices' rejection rates with few and many groups for various combinations of conditions. Tables 8 and 9 give aggregate fit indices' rejection rates.

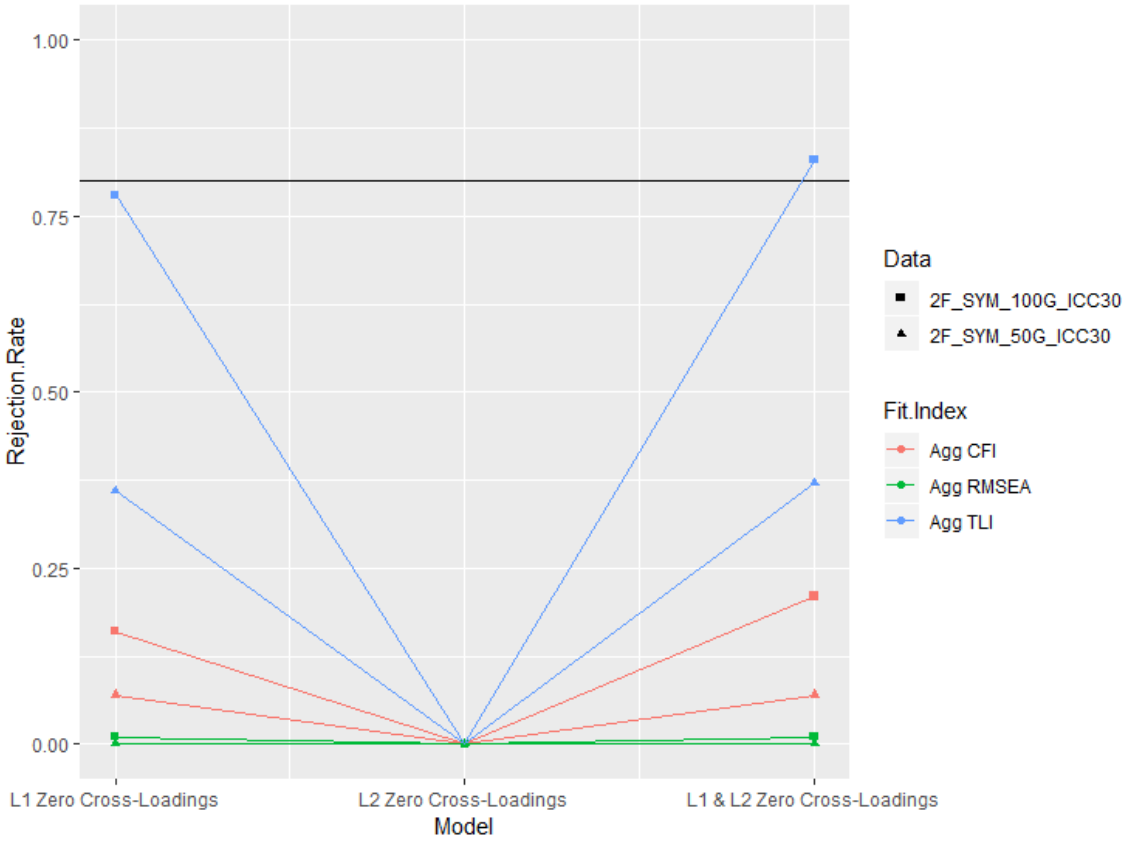

Figure 43. Impact of Number of Groups on Aggregate Fit Indices' Rejection of Cross-Loadings Fixed to 0 with Small Models, Symmetric Data, and Large ICCs.
Note. This figure has overlap. For Level-2 zero cross-loadings, all aggregate fit indices' rejection rates were zero regardless of number of groups. Aggregate RMSEA's rejection rates were identical or nearly identical (.01 difference) for all misspecification conditions for few and many groups.

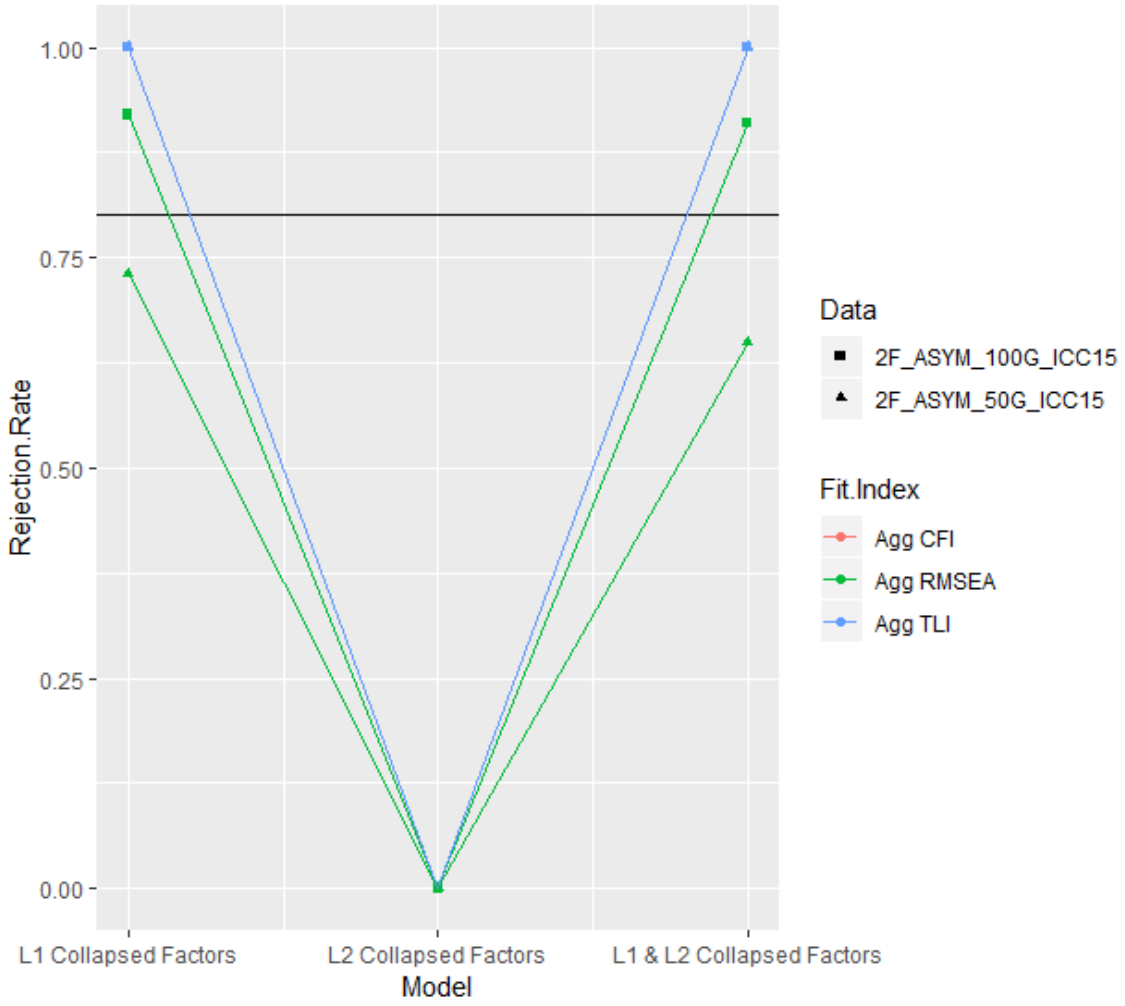

Figure 44. Impact of Number of Groups on Aggregate Fit Indices' Rejection of Collapsed Factors with Small Models, Asymmetry, and Small ICCs.
Note. For collapsed Level-1 factors or both Level-1 and Level-2 collapsed factors. aggregate TLI and CFI had identical rejection rates of 1 regardless of number of groups. For Level-2 collapsed factors, aggregate RMSEA, TLI, and CFI had identical rejection rates of zero regardless of number of groups.

Aggregate fit indices' rejection of Level-1 cross-loadings fixed to 0. Number of groups differentially affected aggregate CFI rejection of Level-1 cross-loadings fixed to 0 based on data symmetry. With asymmetric data, aggregate CFI rejection rates usually decreased as number of groups increased. These decreases ranged from .01 to .07, yielding rejection rates of .02 to .17 with few groups and 0 to .12 with many groups. With large models and symmetric data, aggregate CFI rejection rates were unaffected by number of groups (always .01). With small models and symmetric data, number of groups inconsistently affected CFI rejection rates. For small models, symmetric data, and large ICCs, CFI rejection rates increased from .07 with few groups to .16 with many groups. For small models, symmetric data, and small ICCs, CFI rejection rates decreased from .18 with few groups to .17 with many groups.

With small models, aggregate TLI rejection rates always increased as number of groups increased. Rejection rates were .33 to .63 with few groups and .59 to .85 with many groups. For small models and few groups, TLI performed best with symmetric data and small ICCs (rejection rate = .63). For small models and few groups, aggregate TLI performed worst with asymmetric data and large ICCs (rejection rate = .33). For small models and many groups, TLI performed best with symmetric data and small ICCs (rejection rate = .85). Number of groups most affected aggregate TLI with small models, symmetric data, and large ICCs. In this situation, aggregate TLI rejection rates were .36 with few groups and .78 with many groups.

With large models, aggregate TLI almost never rejected Level-1 cross-loadings fixed to zero regardless of number of groups. For large models and symmetric data,

aggregate TLI rejection rates increased as number of groups increased. Rejection rates were .01 with few groups and .04 to .05 with many groups. For large models and asymmetric data, aggregate TLI rejection rates were .03 to .04 with few groups and .02 to .03 with many groups.

Aggregate RMSEA usually never rejected Level-1 cross-loadings fixed to 0 regardless of number of groups. For small models and symmetric data, aggregate RMSEA rejection rates increased as number of groups increased. Rejection rates were 0 to .08 with few groups and .01 to .10 with many groups.

Aggregate fit indices' rejection of collapsed Level-1 factors. Number of groups differentially affected aggregate CFI rejection of collapsed Level-1 factors based on model size. For small models, aggregate CFI rejection rates were always 1 regardless of number of groups. For large models, aggregate CFI rejection rates usually increased as number of groups increased. Rejection rates were .90 to 1 with few groups and .99 to 1 with many groups. Number of groups most affected aggregate CFI with large models, asymmetric data, and large ICCs; rejection rates were .9 with few groups and 1 with many groups.

Aggregate TLI always rejected 100% of collapsed Level-1 factors regardless of number of groups.

Number of groups differentially affected aggregate RMSEA rejection of collapsed Level-1 factors based on model size. For large models, aggregate RMSEA rejection rates were zero regardless of number of groups. For small models, aggregate RMSEA usually increased as number of groups increased. Rejection rates were .10 to .92 with few groups

and .27 to 1 with many groups. With small models and few groups, aggregate RMSEA performed worst with asymmetric data and large ICCs (rejection rate = .10). For small models and few groups, aggregate RMSEA performed best with symmetric data. In this situation, aggregate RMSEA rejection rates were .92 with large ICCs and 1 with small ICCs. With small models, symmetric data, and small ICCs, aggregate RMSEA rejection rates were 1 regardless of number of groups.

Aggregate fit indices' rejection of Level-2 cross-loadings fixed to 0. Aggregate CFI, TLI, and RMSEA never rejected Level-2 cross-loadings fixed to 0 regardless of number of groups.

Aggregate fit indices' rejection of collapsed Level-2 factors. With one exception, aggregate CFI, TLI, and RMSEA never rejected collapsed Level-2 factors regardless of number of groups. For small models, asymmetric data, and large ICCs, aggregate RMSEA rejection rates were 0 with few groups and .01 with many groups.

Aggregate fit indices' rejection of cross-loadings fixed to 0 at both levels. With large models, aggregate CFI rejection of cross-loadings fixed to 0 at both levels usually was unaffected by number of groups. In these situations, aggregate CFI rejection rates were always .01 regardless of number of groups. With large models, asymmetric data, and small ICCs, aggregate CFI rejection rates were .02 with few groups and 0 with many groups.

With small models, aggregate CFI rejection of cross-loadings fixed to 0 at both levels usually increased as number of groups increased. Rejection rates were .07 to .18 with few groups and .16 to .21 with many groups.

Aggregate TLI rejection of cross-loadings fixed to 0 at both levels usually increased as number of groups increased. For large models, rejection rates were .01 to .03 with few groups and .04 to .05 with many groups. With large models, severe asymmetry, and large ICCs, aggregate TLI rejection rates were .04 regardless of number of groups. For small models, these increases in rejection rates ranged from .22 to .46, producing rejection rates of .34 to .59 with few groups and .63 to .84 with many groups. Number of groups most affected aggregate TLI with small models, symmetric data, and large ICCs. This situation yielded rejection rates of .37 with few groups and .83 with many groups (increase of .46). For small models and few groups, aggregate TLI performed worst with large ICCs (rejection rates of .34 to .37). Aggregate TLI performed best with small models, many groups, symmetric data, and small ICCs (rejection rate = .84).

Aggregate RMSEA usually never rejected cross-loadings fixed to 0 at both levels regardless of number of groups. For small models and symmetric data, RMSEA rejection rates increased as number of groups increased. Rejection rates were 0 to .06 with few groups and .01 to .08 with many groups.

Collapsed factors at both levels. Number of groups differentially affected aggregate CFI rejection of collapsed factors at both levels based on model size. For small models, aggregate CFI rejection rates were always 1 regardless of number of groups. For large models, aggregate CFI rejection rates usually increased as number of groups increased. Rejection rates were .94 to .98 with few groups and .99 to 1 with many groups. For large models and few groups, aggregate CFI performed worst with asymmetric data and large ICCs (rejection rate = .94). Otherwise, large model rejection rates with few

groups were .97 to .98. With large models, symmetric data, and small ICCs, aggregate CFI rejection rates were 1 regardless of number of groups.

Aggregate TLI always rejected 100% of collapsed factors at both levels regardless of number of groups.

Number of groups differentially affected aggregate RMSEA rejection of collapsed factors at both levels based on model size. Aggregate RMSEA never rejected large collapsed models regardless of number of groups. Aggregate RMSEA small model rejection rates usually increased as number of groups increased. Rejection rates were .09 to .90 with few groups and .24 to 1 with many groups. For small models, aggregate RMSEA performed worst with asymmetric data and large ICCs. This situation produced rejection rates of .09 with few groups and .24 with many groups. Aggregate RMSEA performed best with small models fit to symmetric data, yielding rejection rates of .90 to 1 with few groups and 1 with many groups. For small models, symmetric data, and small ICCs, aggregate RMSEA rejection rates were 1 regardless of number of groups.

Correct model at both levels. Aggregate CFI, TLI, and RMSEA never rejected the correct model at both levels regardless of number of groups.

ICCs' Impact on Aggregate Fit Indices

Before describing ICCs' impact on aggregate fit indices' rejection rates, some figures will be shown. These figures introduce the complex nature of aggregate fit index performance. ICCs' impact generally also depended on the specific fit index, type of misspecification, level(s) modeled, number of groups, model size, and data asymmetry.

Figures 45 through 47 contrast aggregate fit indices' rejection rates with small and large ICCs for various conditions. Tables 8 and 9 give aggregate fit indices' rejection rates.

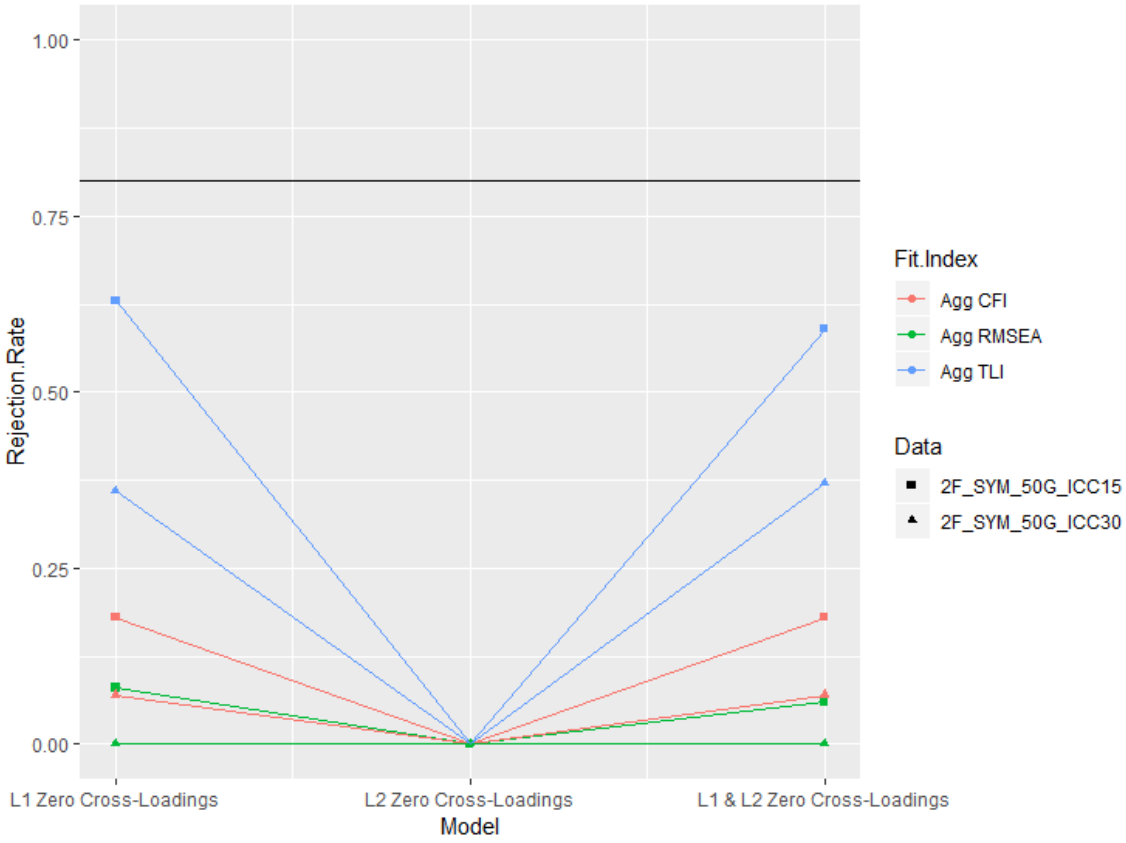

Figure 45. ICC Impact on Aggregate Fit Indices' Rejection of Cross-Loadings Fixed to Zero with Small Models, Symmetry, and Few Groups.
Note. This figure has overlap. For Level-2 zero cross-loadings, all aggregate fit indices' rejection rates were zero regardless of ICCs. For Level-1 cross-loadings or Level-1 and Level-2 cross-loadings, rejection rates differed by .01 for aggregate CFI with large ICCs and aggregate RMSEA with small ICCs.

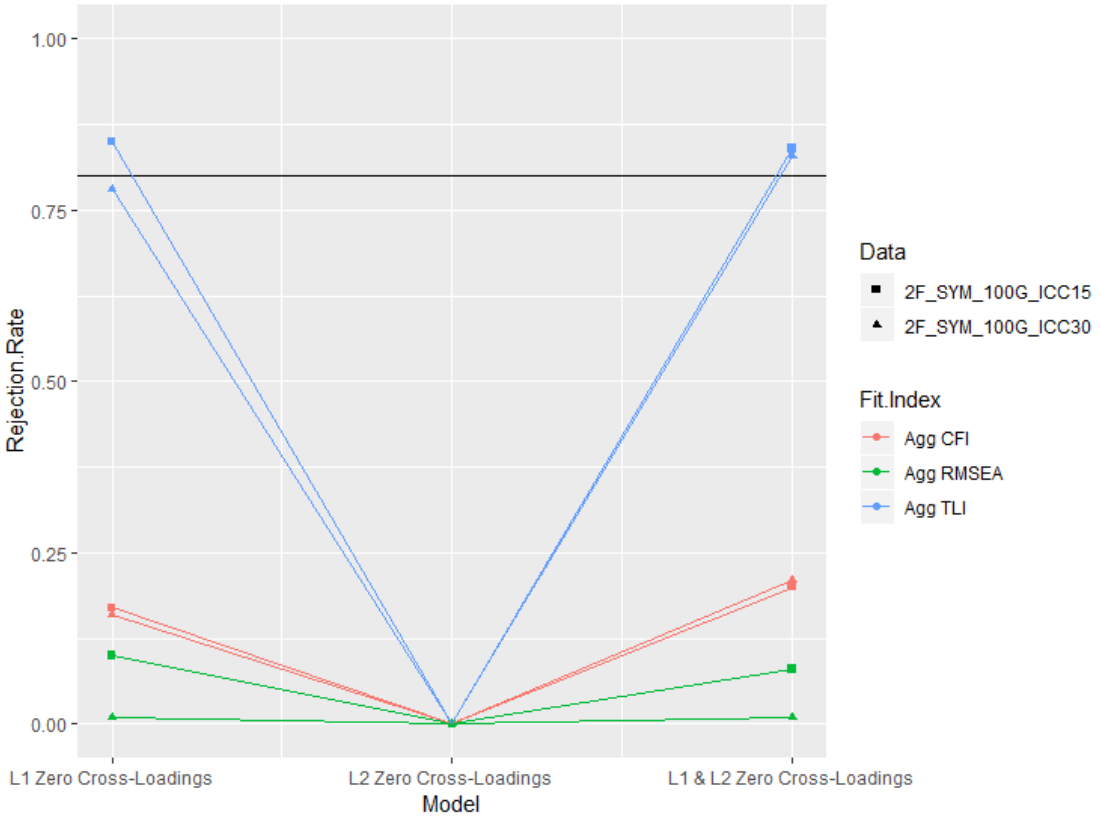

Figure 46. ICC Impact on Aggregate Fit Indices' Rejection of Cross-Loadings Fixed to Zero with Small Models, Symmetric Data, and Many Groups.
Note. This figure has overlap. For Level-2 zero cross-loadings, all aggregate fit indices' rejection rates were zero regardless of ICCs. For Level-1 zero cross-loadings, aggregate CFI rejection rates differed by .01 for small and large ICCs. For Level-1 and Level-2 zero cross-loadings, aggregate CFI rejection rates differed by .01 for small and large ICCs.

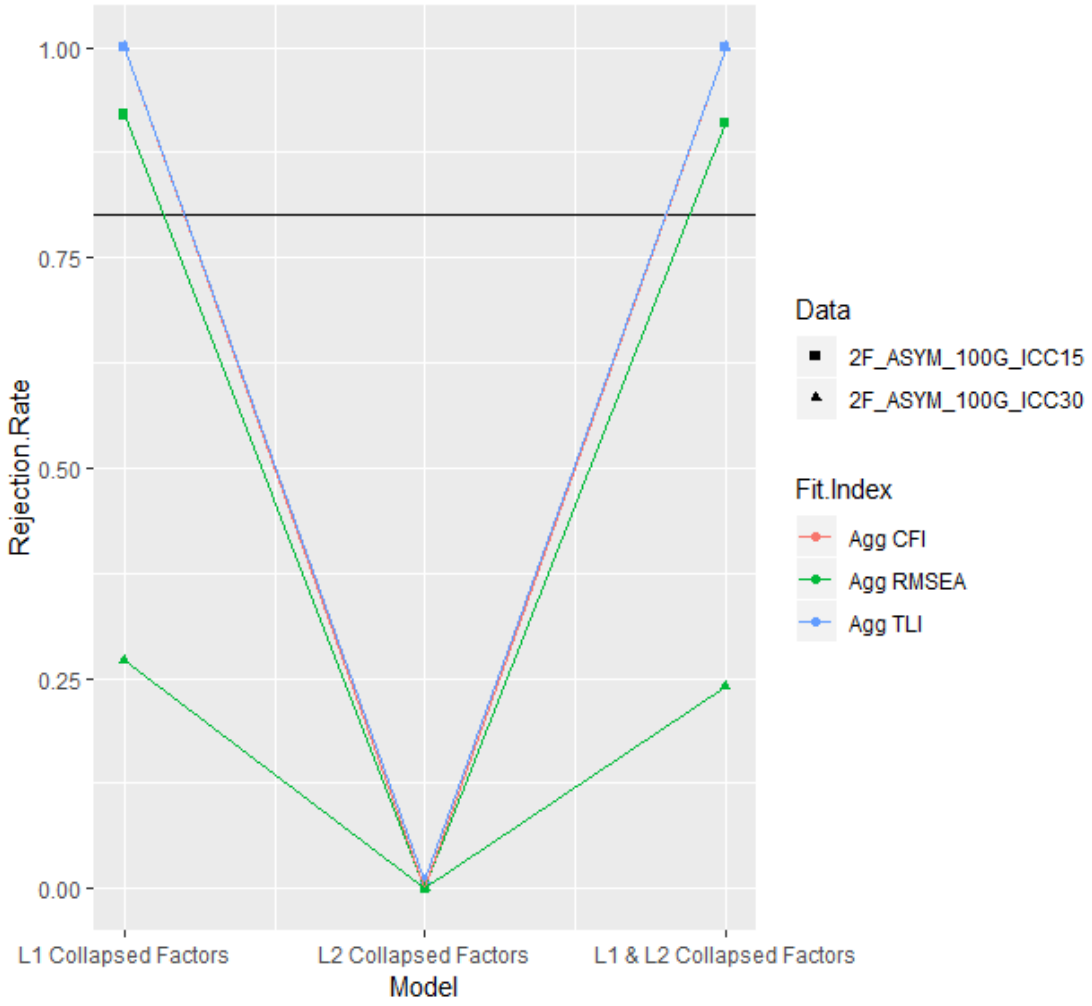

Figure 47. ICC Impact on Aggregate Fit Indices' Rejection of Collapsed Factors with Small Models, Asymmetric Data, and Many Groups.

Note. This figure has overlap. For Level-2 collapsed factors, all aggregate fit indices' rejection rates were zero or .01 regardless of ICCs. Aggregate TLI and CFI's rejection rates were 1 for Level-1 collapsed factors and collapsed factors at both levels regardless of ICCs.

Level-1 cross-loadings fixed to 0. ICCs differentially impacted aggregate CFI rejection of Level-1 cross-loadings fixed to 0 based on model size. With small models, CFI rejection rates usually decreased as ICCs increased. Rejection rates were .17 to .18 with small ICCs and .12 to .16 with large ICCs. With large models, CFI rejection rates generally were close to zero regardless of ICCs. CFI rejection rates for large models were 0 to .02 with small ICCs and .01 to .02 with large ICCs.

ICCs differentially impacted aggregate TLI rejection of Level-1 cross-loadings fixed to 0 based on model size. With small models, TLI rejection rates always decreased as ICCs increased. Small model rejection rates were .43 to .85 with small ICCs and .33 to .78 with large ICCs. For small models, TLI performed worst with severe asymmetry and few groups, yielding rejection rates of .43 with small ICCs and .33 with large ICCs. For small models, TLI performed best with symmetric data and many groups, yielding rejection rates of .85 with small ICCs and .78 with large ICCs. TLI large model rejection rates were .01 to .05 with small ICCs and .01 to .04 with large ICCs.

ICCs differentially impacted aggregate RMSEA rejection of Level-1 cross-loadings fixed to 0 based on model size. With large models, RMSEA rejection rates were always zero regardless of ICCs. With small models and symmetric data, RMSEA rejection rates decreased as ICCs increased, yielding rejection rates of .08 to .10 with small ICCs and 0 to .01 with large ICCs. With small models and asymmetric data, RMSEA rejection rates were always zero regardless of ICCs.

Level-1 collapsed factors. ICCs differentially impacted aggregate CFI rejection of collapsed Level-1 factors based on model size. With small models, CFI always

rejected 100% of collapsed factors regardless of ICCs. With large models, CFI rejection rates usually decreased as ICCs increased. Rejection rates were .97 to 1 with small ICCs and .90 to 1 with large ICCs. With large models, ICCs most affected aggregate CFI with severe asymmetry and few groups, yielding rejection rates of .98 with small ICCs and .90 with large ICCs.

Aggregate TLI always rejected 100% of collapsed Level-1 factors regardless of ICCs.

ICCs differentially impacted aggregate RMSEA rejection of collapsed Level-1 factors based on model size. With large models, RMSEA rejection rates were 0 regardless of ICCs. With small models, RMSEA rejection rates usually decreased as ICCs increased, yielding rejection rates of .73 to 1 with small ICCs and .27 to .92 with large ICCs. For small models, ICCs most negatively impacted RMSEA rejection rates with severely asymmetric data. In these situations, RMSEA rejection rates were .73 to .92 with small ICCs and .10 to .27 with large ICCs. RMSEA always rejected 100% of small collapsed models with symmetric data and many groups regardless of ICCs.

Level-2 cross-loadings fixed to 0. Aggregate CFI, TLI, and RMSEA never rejected Level-2 cross-loadings fixed to 0 regardless of ICCs.

Collapsed Level-2 factors. Aggregate CFI, TLI, and RMSEA almost never rejected collapsed Level-2 factors regardless of ICCs. However, with small models, asymmetry, and many groups, TLI rejection rates were 0 with small ICCs and .01 with large ICCs.

Cross-loadings fixed to 0 at both levels. ICCs differentially affected aggregate CFI rejection of cross-loadings fixed to 0 at both levels depending on model size. For large models, ICCs generally did not affect CFI rejection rates. CFI large model rejection rates were 0 to .02 with small ICCs and .01 with large ICCs. For small models and few groups, CFI rejection rates decreased as ICCs increased, yielding rejection rates of .17 to .18 with small ICCs and .07 to .15 with large ICCs. For small models and many groups, CFI rejection rates increased as ICCs increased. These situations produced rejection rates of .10 to .20 with small ICCs and .16 to .21 with large ICCs.

ICCs differentially affected aggregate TLI rejection of cross-loadings fixed to 0 at both levels depending on model size. For large models and asymmetric data, TLI rejection rates increased as asymmetry increased. Rejection rates were .02 to .03 with small ICCs and .04 with large ICCs. For large models and symmetric data, TLI rejection rates either decreased by .01 or did not change as ICCs increased. TLI rejection rates with large models and symmetric data were .01 to .05 with small ICCs and .01 to .04 with large ICCs.

Aggregate RMSEA usually never rejected cross-loadings fixed to 0 at both levels regardless of ICCs. However, with small models and symmetric data, RMSEA rejection rates decreased as ICCs increased, yielding rejection rates of .06 to .08 with small ICCs and 0 to .01 with large ICCs.

Collapsed factors at both levels. ICCs differentially affected aggregate CFI rejection of collapsed factors at both levels depending on model size. For small models, aggregate CFI rejection rates were always 1.0 regardless of ICCs. For large models,

aggregate CFI rejection rates usually decreased as ICCs increased. Rejection rates were .98 to 1 with small ICCs and .94 to .97 with large ICCs. With large models, symmetric data, and many groups, aggregate CFI rejection rates were always 1 regardless of ICCs. With large models, asymmetric data, and many groups, CFI rejection rates were .99 with small ICCs and 1.0 with large ICCs.

Aggregate TLI always rejected 100% of collapsed factors at both levels regardless of ICCs.

ICCs differentially affected aggregate RMSEA rejection of collapsed factors at both levels depending on model size. For large models, aggregate RMSEA rejection rates were always zero regardless of ICCs. For small models, aggregate RMSEA rejection rates usually decreased as ICCs increased, yielding rejection rates of .65 to 1.0 with small ICCs and .09 to .90 with large ICCs. ICCs most affected RMSEA small model rejection with asymmetric data. With small models and asymmetric data, rejection rates were .65 to .91 with small ICCs and .09 to .24 with large ICCs. The largest decrease occurred with small models, asymmetric data, and many groups. In this situation, rejection rates were .91 with small ICCs and .24 with large ICCs. Finally, with small models, symmetric data, and many groups, aggregate RMSEA rejection rates were always 1 regardless of ICCs.

Correct model at both levels. Aggregate CFI, TLI, and RMSEA never rejected the correct model at both levels regardless of ICCs.

Comparing Level-1, Level-2, and Aggregate Fit Performance

Level-2 fit indices generally rejected Level-2 cross-loadings much more frequently than Level-1 fit indices rejected Level-1 cross-loadings (see Figure 48). Level-

2 fit indices sometimes rejected correct models, whereas Level-1 and aggregate fit indices never rejected correct models (see Figure 49). Level-1 fit indices generally did better than Level-2 fit indices with small ICCs (more Level-1 variance; see Figure 50). Aggregate fit indices often outperformed Level-1 fit indices (see Figures 51 and 52). However, Level-2 fit indices always outperformed aggregate fit indices, as aggregate fit indices never rejected Level-2 misspecified models (see Figures 53 and 54).

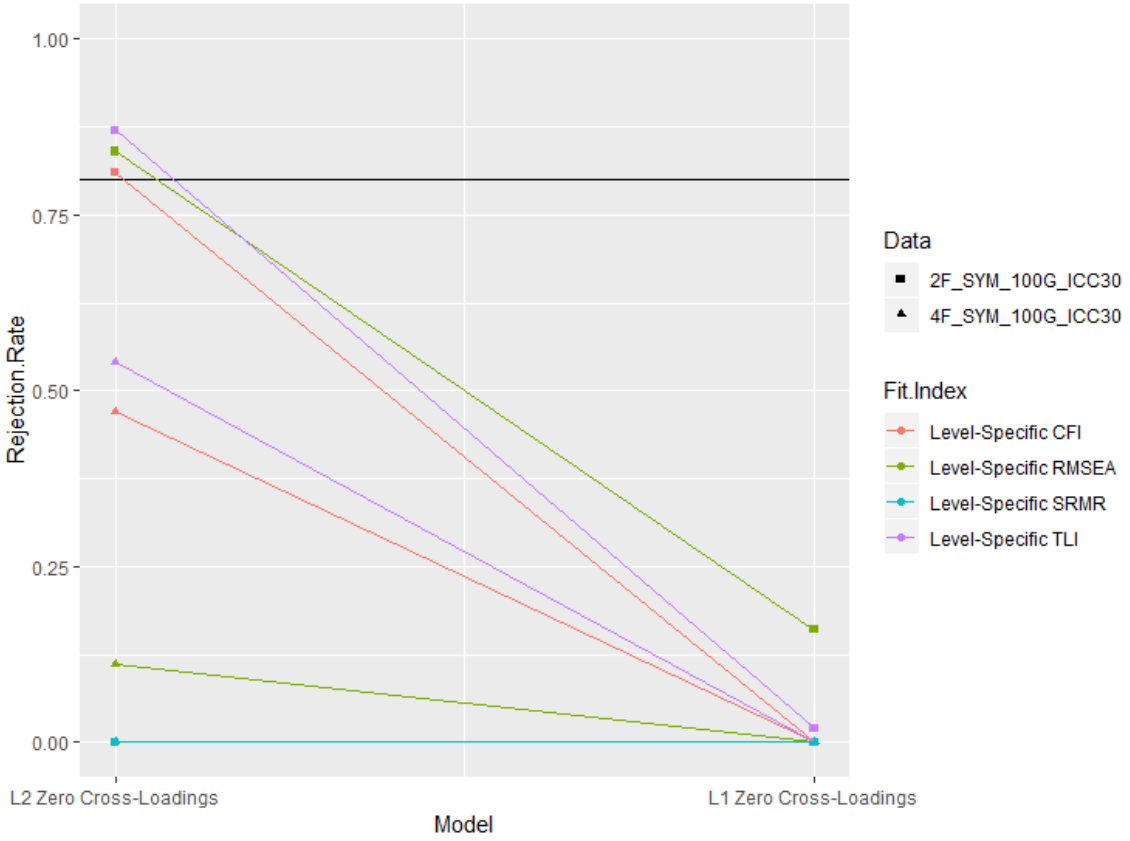

Figure 48. Model Size Impact on Level-Specific Fit Index Performance for Level-1 and Level-2 Cross-Loadings Fixed to 0 with Symmetric Data, Many Groups, and Large ICCs. *Note*. This figure has overlap. Level-specific SRMR always had rejection rates of zero regardless of model size or level of interest. For Level-1 zero cross-loadings, rejection rates were zero for all level-specific fit indices with large models and level-specific SRMR and CFI with small models.

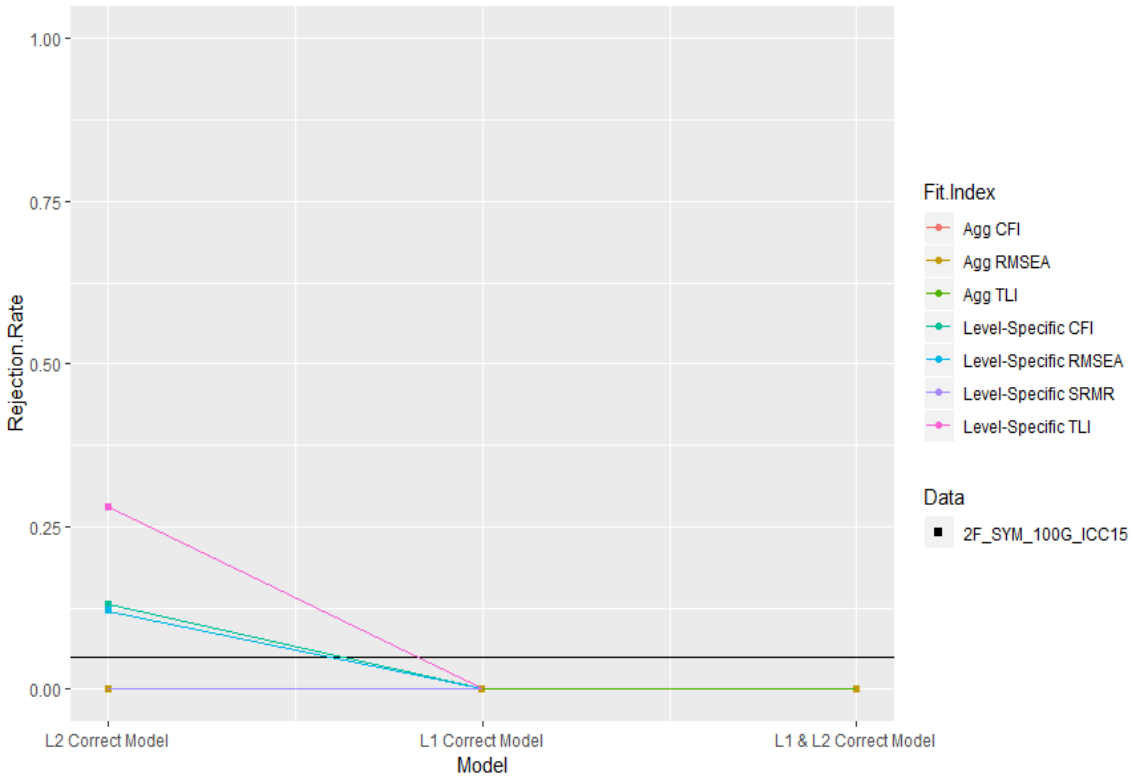

Figure 49. Level-Specific and Aggregate Fit Index Rejection of Correct Model at Level-1 and/or Level-2 with Small Models, Symmetric Data, Many Groups, and Small ICCs. *Note*. This figure has overlap. Level-specific fit indices never rejected correct Level-1 models. Level-specific SRMR was the only level-specific index to never reject correct Level-2 models. Aggregate fit indices never rejected any correct models at either or both levels.

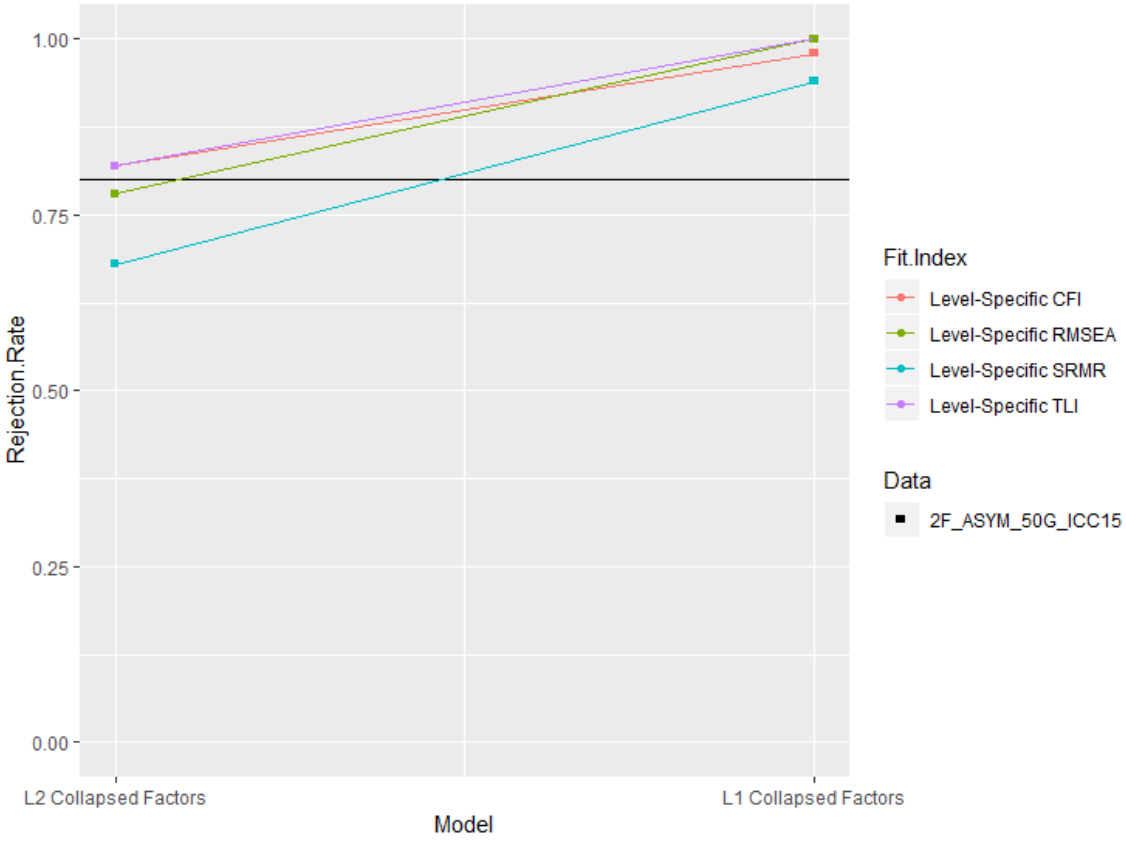

Figure 50. Level-Specific Fit Index Rejection of Collapsed Factors with Small Models, Asymmetric Data, Few Groups, and Small ICCs.
Note. This figure has overlap. Level-specific RMSEA and TLI had identical rejection rates for collapsed Level-1 factors. Level-specific TLI and CFI had identical rejection rates for collapsed Level-2 factors.

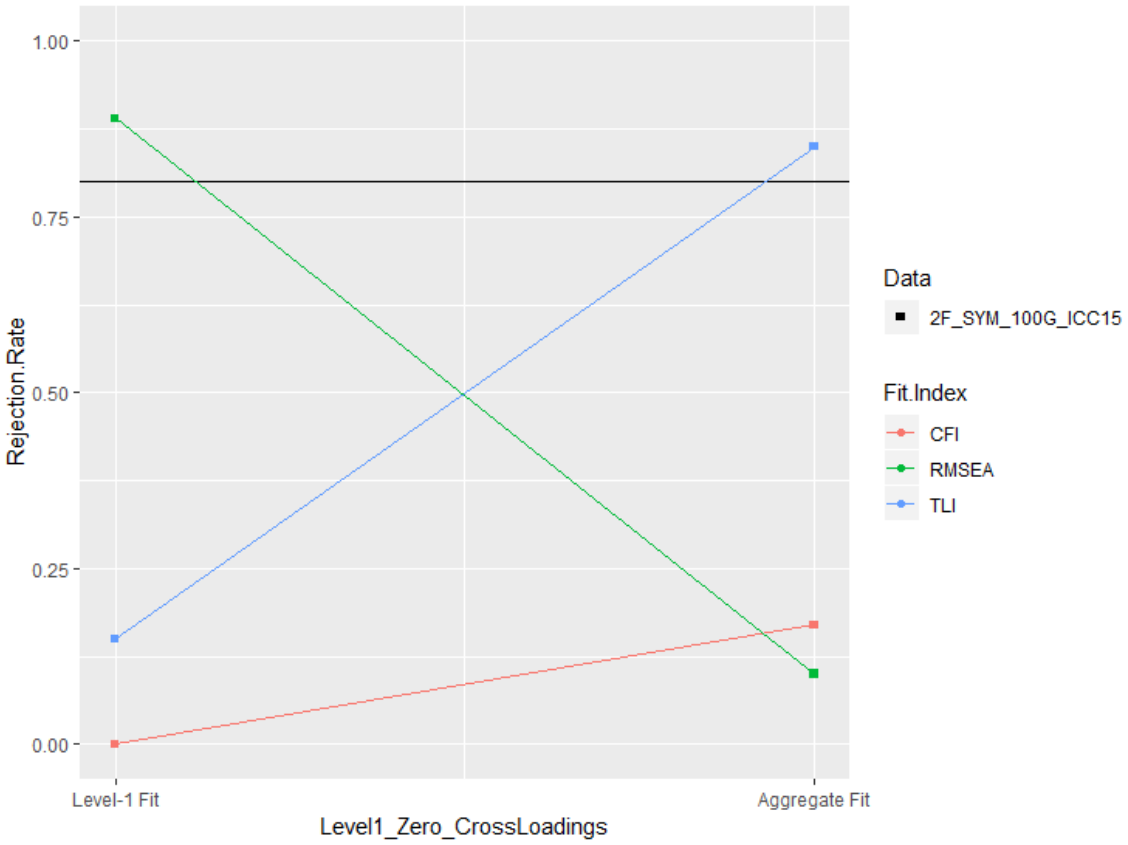

Figure 51. Level-Specific and Aggregate Fit Index Rejection of Level-1 Cross-Loadings Fixed to 0 with Small Models, Symmetric Data, Many Groups, and Small ICCs.

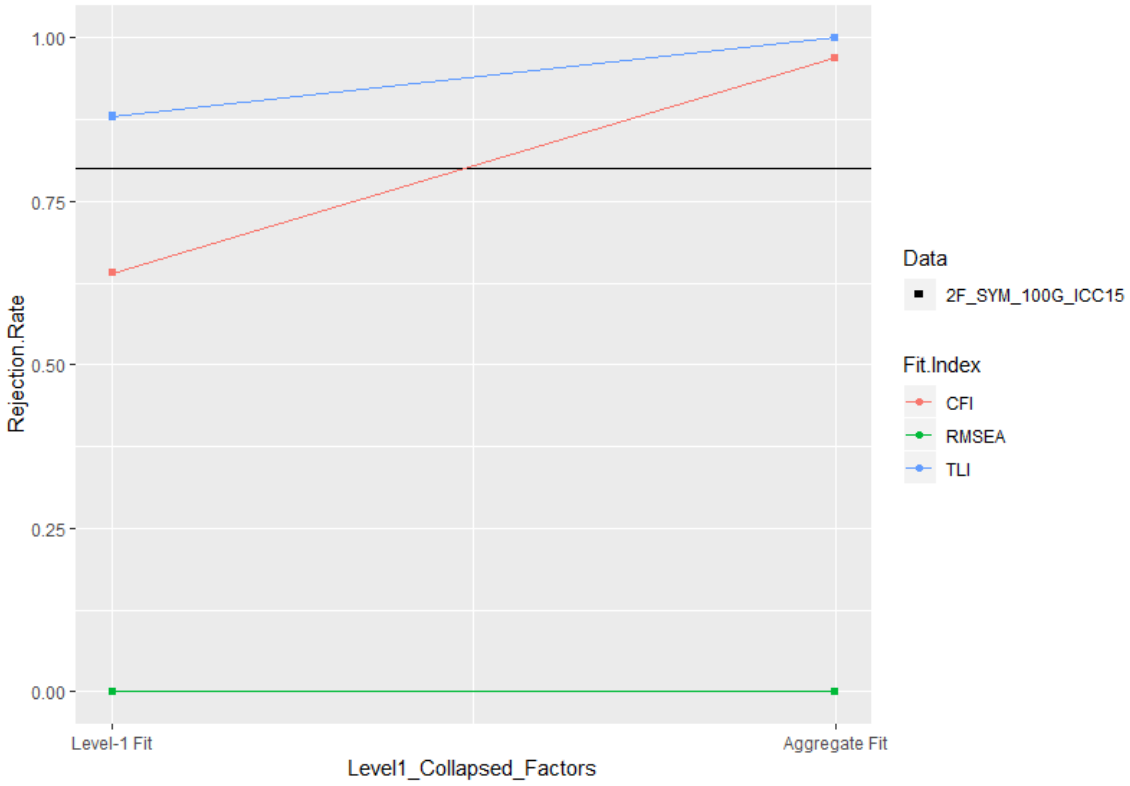

Figure 52. Level-Specific and Aggregate Fit Index Rejection of Level-1 Collapsed Factors with Small Models, Symmetric Data, Many Groups, and Small ICCs.

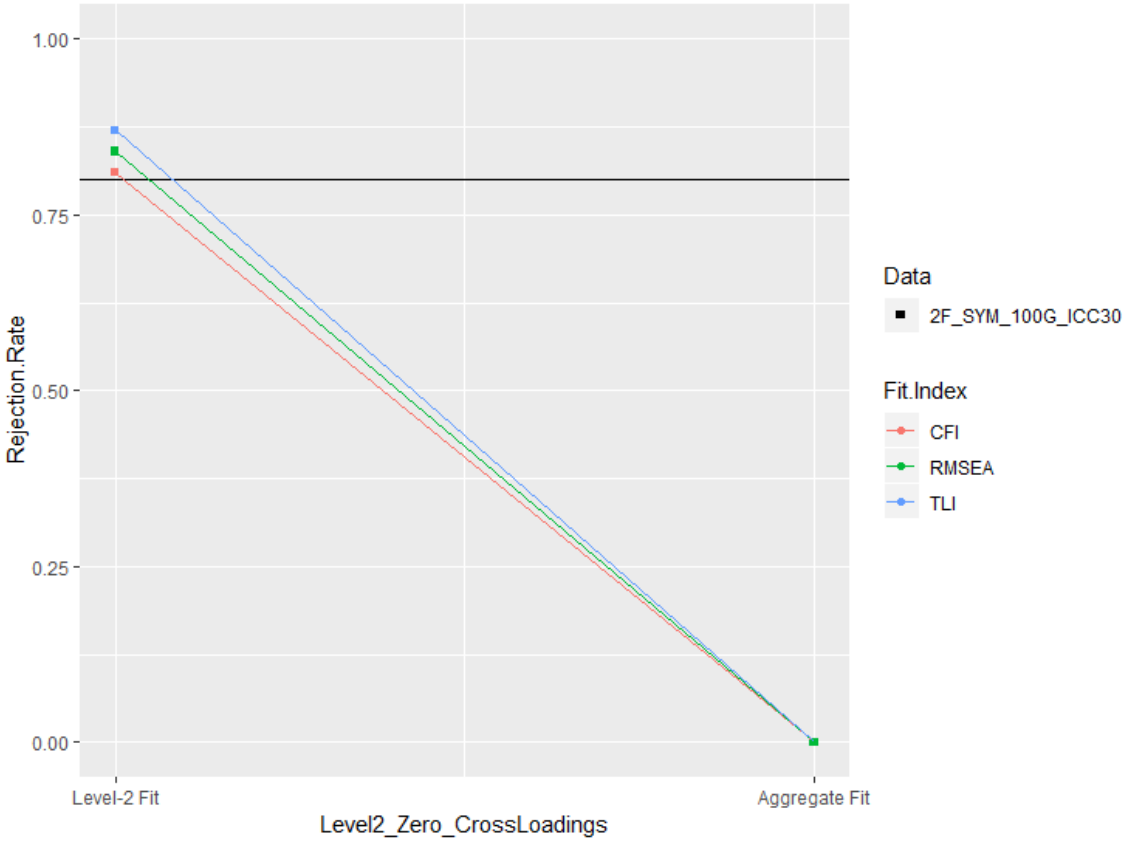

Figure 53. Level-Specific and Aggregate Fit Index Rejection of Level-2 Cross-Loadings Fixed to 0 with Small Models, Symmetric Data, Many Groups, and Large ICCs.
Note. This figure has overlap. Aggregate CFI, TLI, and RMSEA had identical rejection rates of zero for Level-2 crossloadings fixed to zero.

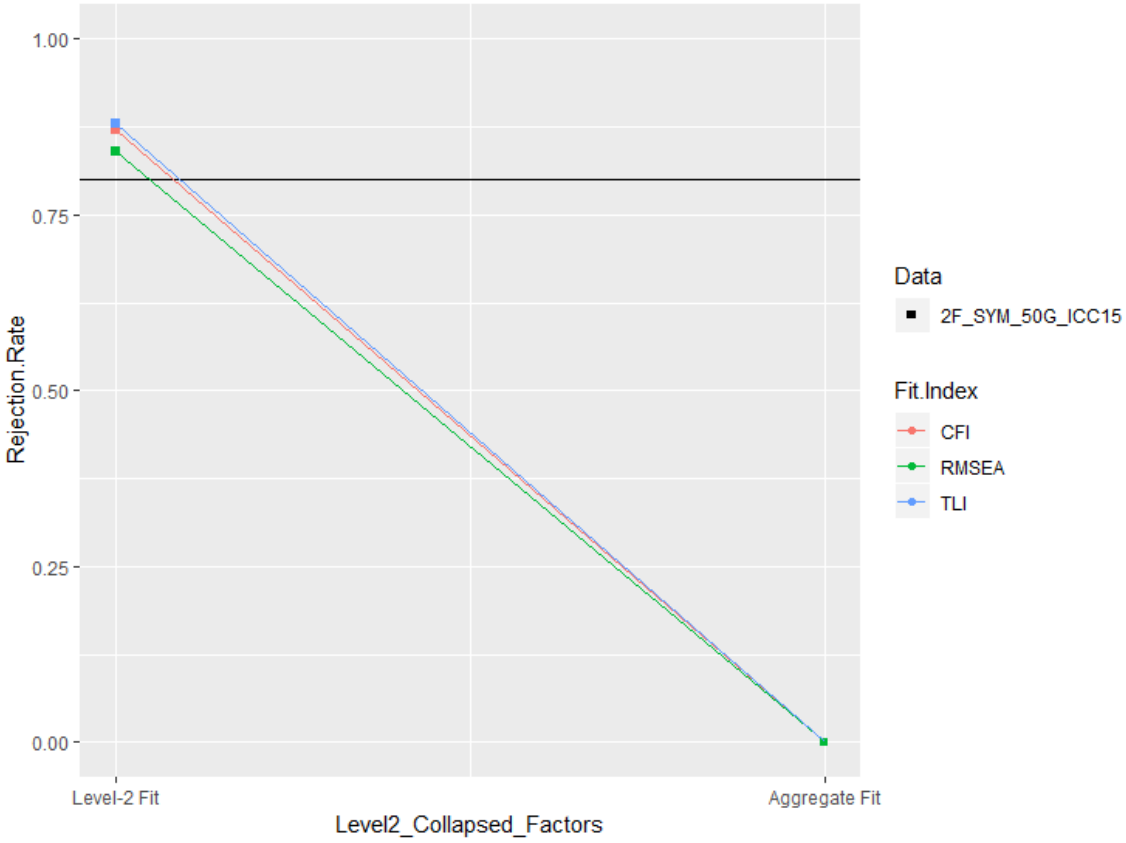

Figure 54. Level-Specific and Aggregate Fit Index Rejection of Level-2 Collapsed Factors with Small Models, Symmetric Data, Few Groups, and Small ICCs.
Note. This figure has overlap. Aggregate CFI, TLI, and RMSEA had identical rejection rates of zero for collapsed Level-2 factors. Level-2 CFI and TLI differed by .01.

Cross-loadings fixed to 0. Level-2 fit indices usually rejected cross-loadings fixed to 0 more frequently than did Level-1 fit indices. Level-1 CFI never rejected Level-1 cross-loadings fixed to 0, whereas Level-2 CFI rejection rates were .16 to .49 for large models and .45 to .81 for small models. Level-2 CFI performed best with small models, symmetric data, many groups, and large ICCs (rejection rate = .81). Level-2 CFI performed worst with large models, symmetric data, few groups, and large ICCs (rejection rate = .16).

Level-1 TLI rejection rates for cross-loadings fixed to 0 were 0 to .15, whereas Level-2 TLI rejection rates were .20 to .57 for large models and .52 to .87 for small models. Level-2 TLI performed best with small models, symmetric data, many groups, and large ICCs (rejection rate = .87). Level-2 TLI performed worst with large models, symmetric data, few groups, and large ICCs (rejection rate = .20).

Level-2 RMSEA usually rejected cross-loadings fixed to 0 more frequently than did Level-1 RMSEA. Level-1 RMSEA rejection rates were 0 for large models and 0 to .89 for small models. Level-2 RMSEA rejection rates were .02 to .06 for large models and .39 to .84 for small models. For small models, Level-2 RMSEA performed worst with asymmetric data, few groups, and small ICCs (rejection rate = .39). For small models, Level-2 RMSEA performed best with symmetric data, many groups, and large ICCs (rejection rate = .84). For small models, Level-1 RMSEA only performed well with symmetric data and small ICCs (rejection rates = .73 to .89). Otherwise, for small models, Level-1 RMSEA rejection rates were 0 to .16.

Aggregate fit and Level-1 fit indices differentially rejected Level-1 cross-loadings fixed to 0 based on model size. Level-1 CFI never rejected Level-1 cross-loadings fixed to 0, whereas aggregate CFI rejection rates were 0 to .17. Level-1 TLI rejection rates were 0 to .15, whereas aggregate TLI rejection rates were 0 to .04 for large models and .33 to 85 for small models. For small models, aggregate TLI performed worst with asymmetric data, few groups, and large ICCs (rejection rate = .33). For small models, aggregate TLI performed best with symmetric data, many groups, and small ICCs (rejection rates = .85). Aggregate RMSEA rejection rates were 0 to .10, whereas Level-1 RMSEA rejection rates were 0 for large models and 0 to .89 for small models. For small models, Level-1 RMSEA only performed well with symmetric data and small ICCs (rejection rates = .73 to .89). Otherwise, for small models, Level-1 RMSEA rejection rates were 0 to .16.

Aggregate CFI never rejected Level-2 cross-loadings fixed to 0, whereas Level-2 CFI rejection rates were .16 to .49 for large models and .45 to .81 for small models. Level-2 CFI performed best with small models, symmetric data, many groups, and large ICCs (rejection rate = .81). Level-2 CFI performed worst with large models, symmetric data, few groups, and large ICCs (rejection rate = .16).

Aggregate TLI never rejected Level-2 cross-loadings fixed to 0, whereas Level-2 TLI rejection rates were .20 to .57 for large models and .52 to .87 for small models. Level-2 TLI performed best with small models, symmetric data, many groups, and large ICCs (rejection rate = .87). Level-2 TLI performed worst with large models, symmetric data, few groups, and large ICCs (rejection rate = .20).

Aggregate RMSEA never rejected Level-2 cross-loadings fixed to 0, whereas Level-2 RMSEA rejection rates were .02 to .06 for large models and .39 to .84 for small models. For small models, Level-2 RMSEA performed worst with asymmetric data, few groups, and small ICCs (rejection rate = .39). For small models, Level-2 RMSEA performed best with symmetric data, many groups, and large ICCs (rejection rate = .84).

Collapsed factors. Level-2 CFI outperformed Level-1 CFI with large collapsed models, whereas Level-1 CFI outperformed Level-2 CFI with small collapsed models. For large collapsed models, Level-1 CFI rejection rates were .44 to .87, whereas Level-2 CFI rejection rates were .75 to 1. For small collapsed models, Level-1 CFI rejection rates were .94 to 1, whereas Level-2 CFI rejection rates were .82 to 1.

Level-2 TLI outperformed Level-1 TLI with large collapsed models, whereas Level-1 TLI outperformed Level-2 TLI with small collapsed models. For large collapsed models, Level-1 TLI rejection rates were .64 to 1, whereas Level-2 TLI rejection rates were .77 to 1. For small collapsed models, Level-1 TLI rejection rates were .98 to 1, whereas Level-2 TLI rejection rates were .82 to 1.

Level-1 RMSEA only rejected large collapsed Level-1 models with symmetric data and small ICCs. These situations yielded Level-1 RMSEA large model rejection rates of .60 with few groups and .97 with many groups.

Level-1 SRMR and Level-2 SRMR rejected collapsed models differently based on model size. Level-1 SRMR rejected small collapsed models more frequently than Level-2 SRMR. Level-2 SRMR rejected large collapsed models more frequently than Level-1 SRMR. For large models, Level-1 SRMR rejection rates were 0 to .04, whereas

Level-2 SRMR rejection rates were .36 to .99. For small models, Level-1 SRMR rejection rates were .54 to .97, whereas Level-2 SRMR rejection rates were .39 to .60. For large models, Level-2 SRMR performed worst with symmetric data, many groups, and large ICCs (rejection rate = .36). For small models, Level-1 SRMR performed worst with symmetric data, many groups, and large ICCs (rejection rate = .54). For large models, Level-2 SRMR performed best with asymmetric data, few groups, and small ICCs (rejection rate = .99). Similarly, for small models, Level-1 SRMR performed best with asymmetric data and small ICCs (rejection rates = .94 to .98).

Aggregate CFI rejected collapsed Level-1 factors more frequently than Level-1 CFI, whereas Level-2 CFI rejected collapsed Level-2 factors more frequently than aggregate CFI. With small Level-1 models, rejection rates were always 1 for aggregate CFI and .94 to 1 for Level-1 CFI. With large Level-1 models, rejection rates were .90 to 1 for aggregate CFI and .44 to 1 for Level-1 CFI. Aggregate CFI never rejected collapsed Level-2 factors, whereas Level-2 CFI rejected 75% to 100% of collapsed Level-2 factors.

Aggregate TLI rejected collapsed Level-1 factors more frequently than Level-1 TLI, whereas Level-2 TLI rejected collapsed Level-2 factors more frequently than aggregate TLI. With small Level-1 models, rejection rates were always 1 for aggregate TLI and .98 to 1 for Level-1 TLI. With large Level-1 models, rejection rates were 1 for aggregate TLI and .64 to 1 for Level-1 TLI. Aggregate TLI never rejected collapsed Level-2 factors, whereas Level-2 TLI rejected 77% to 100% of collapsed Level-2 factors.

Level-1 RMSEA sometimes rejected collapsed Level-1 factors more frequently than aggregate RMSEA. Aggregate RMSEA never rejected large collapsed Level-1

models, whereas Level-1 RMSEA only rejected large collapsed Level-1 models with symmetric data and small ICCs. These situations yielded Level-1 RMSEA large model rejection rates of .60 with few groups and .97 with many groups. With small collapsed Level-1 models, Level-1 RMSEA performed better than aggregate RMSEA. Level-1 RMSEA rejection rates exceeded .90 in all but one condition. Aggregate RMSEA rejection rates exceeded .90 in 5 of 8 conditions. For small models, Level-1 RMSEA performed worst with asymmetric data, few groups, and large ICCs, yielding rejection rates of .65. For small models, Aggregate RMSEA performed worst with asymmetric data and large ICCs; rejection rates were .10 with few groups and .27 with many groups.

Level-2 RMSEA always rejected collapsed Level-2 factors more frequently than aggregate RMSEA. Aggregate RMSEA never rejected collapsed Level-2 factors, whereas Level-2 RMSEA rejection rates were .24 to 1. Level-2 RMSEA performed worst with large models and asymmetric data, yielding rejection rates of .24 to .45. Level-2 RMSEA performed best with small models and symmetric data, yielding rejection rates of .84 to 1.

Correct model. Aggregate and Level-1 fit indices never rejected the correct model, whereas Level-2 fit indices often rejected the correct model. Level-2 CFI rejected 1% to 22% of large correct models and 3% to 18% of small correct models. For large models, Level-2 CFI performed worst with asymmetric data, few groups, and small ICCs (rejection rate = .20 to .22). For small models, Level-2 CFI performed worst with symmetric data, few groups, and small ICCs (rejection rate = .18). For small models, Level-2 CFI performed best with symmetric data, many groups, and large ICCs (rejection

rate = .03). For large models, Level-2 CFI performed best with asymmetric data, many groups, and large ICCs (rejection rate = .01).

Level-2 TLI rejection rates for correct models were .01 to .20 for large models and .12 to .28 for small models. Of the Level-2 fit indices, Level-2 TLI performed worst with small correct models. Level-2 TLI performed worst with small models, symmetric data, many groups, and small ICCs (rejection rate = .28). For small models, Level-2 TLI performed best with symmetric data, many groups, and large ICCs (rejection rate = .12). For large models, Level-2 TLI performed best with symmetric data, few groups, and large ICCs (rejection rate = .01).

Level-2 RMSEA rejected 0% to 1% of large correct models and 6% to 20% of small correct models. Although Level-2 RMSEA yielded excellent rejection rates with large correct models, Level-2 RMSEA rejection rates for large misspecified models rarely exceeded .50 (i.e., low power to detect misfit).

Level-2 SRMR was inconsistent with large correct models. Level-2 SRMR usually rejected 0-1% of large correct models. For large correct models, few groups, and small ICCs, Level-2 SRMR rejection rates were .34 for symmetric data and .80 for asymmetric data. With large models, asymmetric data, few groups, and small ICCs, Level-2 SRMR rejection rates exceeded .80 for all misspecified Level-2 models and correct Level-2 models. That is, with the least optimal conditions for Level-2 models, Level-2 SRMR tended to reject almost all models regardless of whether there was no misfit, minor misfit, or major misfit.

Summary of Analysis Results

The analysis results were complex. Model size, model misspecification, number of groups, ICCs, data asymmetry, and level of misspecification each affected fit indices' rejection rates. These conditions often interacted with each other so that their effect varied based on the other conditions. For example, Level-2 RMSEA rejected collapsed factors much more frequently for small models than large models. For large models, Level-2 RMSEA rejected collapsed factors much more frequently with large ICCs than small ICCs. Generally, rejection rates increased as model size decreased, data asymmetry decreased, number of groups increased, and degree of misspecification increased. For Level-2 fit indices, rejection rates generally increased as ICCs increased. For Level-1 fit indices, rejection rates often decreased as ICCs increased. Generally, model size and collapsed factors had the largest and most consistent effects on fit index performance.

Level-2 fit indices' performance was more nuanced than Level-1 and aggregate fit indices' performance. For cross-loadings fixed to 0, Level-2 fit indices' power usually was low (< .5) with large models and medium (> .5) with small models. For collapsed factors, Level-2 CFI and TLI usually had high power (> .80). For collapsed factors, Level-2 RMSEA usually had medium power (~ .5) for large models and high power (> .80) for small models. Level-2 SRMR generally had medium power (~ .5) for small collapsed models and medium to high power (.5 to .9) for large collapsed models. Level-2 CFI, TLI, and RMSEA typically did best in the expected optimal set of conditions (small model, symmetric data, many groups, large ICCs). Level-2 SRMR usually performed much differently than Level-2 CFI, TLI, and RMSEA. Level-2 SRMR usually

had high rejection rates for correct and incorrect models in the least optimal conditions (large model, asymmetric data, few groups, and small ICCs). Level-2 CFI, TLI, and RMSEA typically performed best with small models. Level-2 SRMR usually had high rejection rates with large models (correct and incorrect). Generally, Level-2 CFI and TLI performed best, followed by Level-2 RMSEA, then Level-2 SRMR.

Level-1 fit indices and aggregate fit indices never rejected correct models and almost never rejected Level-1 cross-loadings fixed to 0. Level-1 RMSEA had adequate power (.73 to .89) to reject Level-1 cross-loadings fixed to 0 only with small models, symmetric data, and small ICCs. Aggregate TLI had power of .80 to detect Level-1 cross-loadings fixed to 0 with small models and many groups. Aggregate fit indices generally had strong power (.90 to 1) to reject Level-1 collapsed factor models or Level-1 and Level-2 collapsed models. Aggregate fit indices never rejected Level-2 misspecified models.

Level-2 CFI and TLI usually performed best, then Level-2 RMSEA, then Level-2 SRMR. Generally, Level-1 fit indices performed similarly. However, with collapsed factors, Level-1 TLI performed best, followed by Level-1 CFI. Level-1 SRMR and RMSEA performed well only with small collapsed models. Sometimes these performance differences were very large. With large models, symmetric data, few groups, and large ICCs, Level-1 TLI rejection rates were .82, compared to .49 for Level-1 CFI and 0 for Level-1 RMSEA and SRMR. Among aggregate fit indices, aggregate TLI performed best. Sometimes these performance differences were very large. With small

models, symmetry, many groups, and small ICCs, rejection rates were .85 for aggregate TLI, .13 for aggregate CFI, and .10 for aggregate RMSEA.

CHAPTER V

DISCUSSION

Study Context and Design

This study assessed level-specific fit index performance for Diagonally Weighted Least Squares (DWLS) estimation of multilevel structural equation models (MSEMs). Model fit is used to assess if parameter estimates are accurate and trustworthy. Good model fit suggests that parameter estimates are appropriate to interpret; poor model fit suggests that parameter estimates should not be inspected. MSEMs provide parameter estimates at each level, but applied researchers usually evaluate MSEM fit aggregating across both levels and rarely inspect level-specific fit (Kim et al., 2016). Researchers typically inspect aggregate fit indices because they are provided in the output. Level-specific indices must be computed "by hand". Assessing aggregate fit indices in MSEM is problematic for two reasons. First, misfit can occur at the within level, between level, or both. Assessing aggregate fit can mask at which level(s) the misfit is occurring. Second, Level-1 sample sizes often are much larger than Level-2 samples (e.g. 1000 students and 50 teachers). Fit indices are based on model chi-square, which is influenced heavily by sample size. The larger the sample size, the larger the chi-square; large chi-square values suggest poor fit. With large samples, chi-square often will yield large values (suggesting substantial misfit) even with minor misfit. Because aggregate fit

indices combine (mis)fit at both levels, aggregate fit indices likely primarily reflect Level-1 (mis)fit given a much larger Level-1 sample (Hsu et al., 2015, 2017).

Level-specific fit evaluation assesses model fit separately at each level. Level-specific fit indices usually outperform aggregate fit indices (Boulton, 2011; Ryu & West, 2009; Yuan & Bentler, 2007). Level-specific fit indices are an important contribution to the MSEM literature. The simulation studies that have evaluated level-specific fit indices have contributed substantial insight into MSEMs. However, these simulation studies do share some important limitations that motivated my study. First, they all used Maximum Likelihood estimation and continuous observed variables. In single-level SEM, fit index performance can vary based on estimation method and whether continuous or categorical variables are analyzed (Yu & Muthen, 2002). Categorical variables are common in MSEM applications (Kim et al, 2016), and DWLS estimation usually is advised over ML for categorical MSEM (Depaoli & Clifton, 2015). Second, these MSEM simulation studies all used small models (1-2 factors at each level). Single-level SEM simulations suggest that increasing model size often negatively impacts DWLS power, fit indices, and parameter estimation accuracy (Bandalos, 2014). MSEM applications often use large models with at least 4 factors per level (Kim et al., 2016). MSEM simulation studies are used to inform MSEM applications (e.g. fit index values that indicate good fit). Thus, using model sizes that align with MSEM applications is important. Third, these MSEM simulation studies all used normally distributed variables. Little MSEM simulation research has evaluated non-normality's effects. Single-level DWLS fit indices usually perform increasingly worse as non-normality increases (Bandalos, 2014; Li, 2016).

Severe non-normality is studied much less than mild or moderate non-normality, but also affects DWLS much more (Bandalos, 2014; Li, 2016). Thus, my study incorporated severe non-normality through highly skewed item thresholds.

My simulation design sought to address these limitations and incorporate other aspects that typically affect level-specific fit indices. My study used categorical variables with four response categories and DWLS estimation. Level-1, Level-2, and aggregate fit indices' performance was evaluated. Level-specific indices evaluated were CFI, TLI, RMSEA, and SRMR. Aggregate fit indices inspected were CFI, TLI, and RMSEA. These fit indices are predominant in MSEM simulation studies (e.g., Hsu et al., 2015, 2017). For aggregate fit indices, I only evaluated their performance when fitting the full MSEM (factor model at both levels). Study conditions were model size (small and large), number of groups (few and many), ICCs (small and large), data nonnormality (normal and severely non-normal), type of misspecification (correct model, cross-loadings fixed to zero, or collapsed factors), and level of misspecification (Level-1, Level-2, and both). Because categorical variables cannot be normally distributed, throughout the document I referred to normality and nonnormality as symmetric and asymmetric data (*cf* Bandalos, 2014). Asymmetry/non-nonnormality was manipulated using item thresholds. Number of groups (Level-2 sample size) and ICCs (proportion of Level-2 variance) usually affect level-specific fit performance. As number of groups and ICCs increase, level-specific fit indices usually increase in accuracy (Boulton, 2011; Hsu et al., 2017). Condition values were based on typical values in MSEM simulations and applications to increase generalizability.

Defining Fit Index Performance

Fit index performance was defined using the percent/proportion of replications that indicated poor fit using Hu and Bentler's (1999) suggested good fit values. For example, Hu and Bentler (1999) suggest that CFI values ≥ .95 indicate good model fit. Of the 100 replications, if 75 CFI values were below .95, then 75 models would be rejected as poorly fitting based on the CFI. The rejection rate would be .75 (75/100). Throughout the results I referred to percentages (e.g. rejected 75% of misspecified models) and proportions (e.g. rejection rate of .75). For a given index (e.g. CFI), the same fit cut-off values were applied to Level-1, Level-2, and aggregate fit. For example, the same cut-off value of ≥ .95 was applied to Level-1 CFI, Level-2 CFI, and aggregate CFI.

Overall Results Summary

All study conditions (misspecification, data non-normality, number of groups, ICCs, and model size) affected Level-2 fit index rejection rates. These effects often were inter-dependent but were less consistent for Level-1 and aggregate indices. Within a given type of fit index (e.g. Level-2), fit index performance varied widely across indices (e.g., Level-2 CFI, Level-2 TLI, Level-2 RMSEA, Level-2 SRMR). Rejection rates were usually higher for collapsed factors (large misspecification) than cross-loadings fixed to zero (small misspecification) or correct models (no misspecification). Cross-loadings fixed to zero often had low rejection rates. The unstandardized cross-loading values used perhaps were too small to represent detectable misfit.

Conditions often affected Level-2 fit indices more consistently than Level-1 or aggregate fit indices. Level-1 and aggregate fit indices never rejected correct models

regardless of other conditions. Increasing number of groups usually increased fit indices' rejection of incorrect models. Increasing number of groups and model size typically decreased Level-2 fit indices' rejection of correct models. Increasing model size usually decreased fit index rejection of incorrect models. Increasing non-normality typically decreased rejection of incorrect models. Increasing non-normality usually increased Level-2 fit index rejection of correct models. Increasing ICCs often decreased Level-1 indices' rejection of incorrect models and increased Level-2 indices' rejection of incorrect models. TLI usually performed best, followed by CFI, RMSEA, and SRMR. Level-1, Level-2, and aggregate fit indices' performance will be described in isolation. Then, I will compare Level-1, Level-2, and aggregate fit indices' performance.

Level-1 Fit Indices' Performance

Level-1 fit indices usually had no power to reject Level-1 cross-loadings fixed to zero. Level-1 fit indices' rejection rates were always zero for large models and usually very low for small models (≤ .16). The only exception occurred for Level-1 RMSEA with small models, symmetric data, and small ICCs. This situation yielded rejection rates of .73 with few groups and .89 with many groups.

For Level-1 collapsed factors, only Level-1 CFI and TLI had excellent power (> .8) in most situations. Level-1 TLI performed best, followed by Level-1 CFI, Level-1 RMSEA, and Level-1 SRMR. For small models, Level-1 CFI, TLI, and RMSEA rejection rates usually exceeded .90 (excellent power). For small models, Level-1 SRMR rejection rates always exceeded .50 (medium power). For small models, Level-1 SRMR rarely yielded excellent power (> .80) except with small ICCs (rejection rates = .92 to 1).

For large models, Level-1 TLI rejection rates often were much higher than other Level-1 fit indices. For example, with large models, asymmetric data, many groups, and large ICCs, Level-1 TLI had excellent power (.85), Level-1 CFI had medium power (.52), and Level-1 RMSEA and SRMR had zero power. For large models, Level-1 TLI's power was almost always excellent (.8 to 1). Level-1 CFI had excellent power (.8 to 1) with symmetric data but usually medium power (around .5) with asymmetric data. Level-1 RMSEA usually never rejected large collapsed models (power of zero in most situations). For large models, Level-1 RMSEA only had medium (> .5) or high (> .8) power with symmetric data and small ICCs. Level-1 SRMR never rejected large collapsed models.

Level-1 fit indices never rejected the correct Level-1 model (i.e., no Type I error).

Level-2 Fit Indices' Performance

Model size affected Level-2 fit indices' rejection of Level-2 cross-loadings fixed to zero. Level-2 CFI, TLI, and RMSEA usually yielded medium power (≥ .50) with small models. Level-2 TLI's power approached .80 with small models and many groups. Level-2 CFI, TLI and RMSEA yielded excellent power (≥ .80) only with small models, symmetric data, many groups, and large ICCs. Level-2 CFI (< .50), TLI (< .50), and RMSEA (≤ .10) generally had low power with large models. Level-2 TLI sometimes yielded medium power (≥ .50) with large models. Level-2 SRMR usually yielded very low power (≤ .3) for small models and low power (< .4) for large models. Level-2 SRMR yielded medium to excellent power with large models, few groups, and small ICCs.

Level-2 fit indices' rejection of Level-2 collapsed factors depended on model size. Level-2 CFI and TLI usually yielded excellent power (≥ .80) regardless of model size.

Level-2 RMSEA generally yielded excellent power (≥ .80) for small models. With large models, Level-2 RMSEA usually had low to medium power (.24 to .5). Level-2 RMSEA yielded excellent power (≥ .80) with many groups and large ICCs. Level-2 SRMR generally had high rejection rates when other indices did not. For large models, Level-2 SRMR had high rejection rates (≥ .80) for correct and incorrect models with asymmetry, few groups, and small ICCs. Usually, severe Type I error rates accompanied high power conditions for Level-2 SRMR. Otherwise, Level-2 SRMR had low to medium power (rejection rate = .36 to .68).

For correct Level-2 models, Level-2 SRMR usually performed best, and Level-2 TLI performed worst. Level-2 SRMR rejection rates were 0 except with large models, few groups, and small ICCs (rejection rate = .34 to .8). Level-2 RMSEA always yielded desired rejection rates ≤ .05 for large correct models. Level-2 CFI only provided acceptable Type I error rates (< .05) with many groups and large ICCs. Level-2 TLI only yielded desired rejection rates ≤ .05 with large models, many groups, and large ICCs. Otherwise, Level-2 CFI, TLI, and RMSEA often rejected at least 10% or 20% of correct Level-2 models.

Aggregate Fit Indices' Performance

Aggregate fit indices usually had low power to reject Level-1 cross-loadings fixed to zero. Aggregate CFI's power was very low (< .20) for small models and close to zero (≤ .02) for large models. Aggregate RMSEA's power was very low (≤ .11) for small models and zero for large models. Aggregate TLI often had medium (> .5) or high (> .8)

power for small models and close to zero (≤ .05) for large models. For small models, aggregate TLI had excellent power (≥.78) with symmetric data and many groups.

Aggregate CFI and TLI always yielded excellent power (> .9) with collapsed Level-1 factors. Aggregate RMSEA usually yielded excellent power (> .8) with small collapsed Level-1 models. For small collapsed Level-1 models, aggregate RMSEA had very weak power (.1 to .27) with severely non-normal data and large ICCs. Aggregate RMSEA always had zero power for large collapsed Level-1 models.

Aggregate fit indices never rejected models with only Level-2 misfit or the correct model at both levels. Aggregate fit indices had zero power to detect Level-2 misfit except when Level-1 misfit also occurred. Even then, power levels were similar for misspecification at both levels and Level-1 misspecification only.

Power to reject cross-loadings fixed to zero at both levels depended on fit index and model size. For small models, aggregate RMSEA (≤ .10) and aggregate CFI (≤ .21) had low power. For small models, aggregate TLI often had medium power (≥ .5) or high power (≥.80). Aggregate TLI had excellent power (≥.80) with small models, symmetric data, and many groups. Aggregate CFI, TLI, and RMSEA had zero or close to zero (≤ .05) power for large models.

Aggregate CFI and TLI yielded excellent power (≥ .94) to reject collapsed factors at both levels. Aggregate RMSEA had excellent power (≥ .90) for most small models but yielded very low power (≤ .24) with asymmetric data and large ICCs. Aggregate RMSEA had zero power with large models.

Comparing Level-1, Level-2, and Aggregate Fit Performance

Level-2 fit indices' performance usually varied more than Level-1 fit indices and aggregate fit indices. Level-1 fit indices' rejection rates usually increased as ICCs decreased. Level-2 fit indices' rejection rates usually decreased as ICCs decreased. Level-1 fit indices and aggregate fit indices never rejected the correct model. Level-2 fit indices rejected 10% to 30% of correct models except with very optimal data. Aggregate fit indices never rejected models with only Level-2 misfit, but Level-2 fit indices often rejected Level-2 misfit. Aggregate fit indices' rejection rates for Level-1 misfit only were similar or better than Level-1 fit indices.

Comparing Study Results to the Literature

DWLS level-specific fit performance usually deteriorated as model size and data asymmetry (non-normality) increased. Single-level DWLS performance usually worsens as asymmetry and model size increases (Bandalos, 2008, 2014). DWLS fit indices became increasingly insensitive to misfit as asymmetry increased (Bandalos, 2008). Thus, my findings regarding the impact of model size and data asymmetry for DWLS align with previous single-level SEM research.

Of the level-specific fit indices, level-specific CFI and TLI performed best. Level-specific CFI and TLI previously outperformed level-specific RMSEA and SRMR with continuous variables and ML estimation (Hsu et al., 2017). Previously, Level-2 RMSEA and SRMR performed poorly with small ICCs and collapsed factors (Hsu et al., 2017). My results were similar, but Level-2 SRMR usually only had high power ($\geq .8$) with

some large models and low power (≤ .30) with most small models. Previous simulations only used small models, thus my findings support previous research.

Aggregate CFI, TLI, and RMSEA never rejected misspecified Level-2 models. Aggregate CFI, TLI, and RMSEA did not detect Level-2 misfit in previous MSEM simulations (Boulton, 2011; Hsu et al., 2015; Ryu & West, 2009). The aggregate goodness of fit index (GFI) and adjusted GFI have shown sensitivity to Level-2 misfit (Boulton, 2011). Aggregate GFI and adjusted GFI's ability to detect Level-2 misfit usually increases as ICCs increase (Boulton, 2011).

Aggregate and level-specific fit indices always rejected collapsed factors more frequently than cross-loadings fixed to 0. Previously, level-specific SRMR also rejected collapsed factors more frequently than cross-loadings fixed to 0 (Hsu et al., 2015). However, aggregate fit indices rejected cross-loadings fixed to 0 more than collapsed factors (Hsu et al., 2015). These differences are difficult to explain because collapsed factors represent more total misfit than cross-loadings fixed to zero.

Aggregate fit indices generally outperformed Level-1 fit indices. Aggregate fit indices previously have detected Level-1 misfit about as well as level-specific fit indices (Hsu et al., 2015; Ryu & West, 2009). In my study, aggregate fit indices and level-specific fit indices performed very similarly with small models and symmetric data. These conditions mirror typical MSEM simulation studies that use small models and normally distributed data. The largest differences in performance between aggregate and Level-1 fit indices occurred in the conditions that had not been explored in MSEM (large

models and asymmetric data). Thus, my results regarding aggregate fit compared to Level-1 fit indices align with previous findings.

Implications for Practice

Level-specific TLI usually outperformed or performed as well as other level-specific indices. Thus, when conducting MSEM, practitioners should consult level-specific TLI. Aggregate TLI should be inspected when fitting the full MSEM. Aggregate TLI usually outperformed Level-1 TLI regarding identification of Leve1-1 misfit.

Level-2 SRMR was very inconsistent and could be unusable in many situations. With the worst set of conditions, Level-2 SRMR rejected 80% of correct models. Level-2 SRMR often had high rejection rates for correct and incorrect models with poor data. Thus, despite excellent power with poor data that outperformed other indices, Level-2 SRMR's use is cautioned. Hsu et al. (2017) advised against Level-2 SRMR with small ICCs. Level-2 SRMR's rejection was very sensitive to number of groups. Rejection rates usually greatly increased when number of groups decreased. This result suggests that the correlations that SRMR is based on (i.e., mean correlation residual) are very unstable (estimated poorly) with small Level-2 samples and reproduced inaccurately by the model. This model instability was evident especially with large models and skewed data. Rejection rates were near zero with many groups and close to .90 with few groups. These results suggest many large residuals and a MSEM limitation, where fit and parameters may be unstable with poor data (common in application). Skewness seems to inflate residuals and increase rejection rates and model instability regardless of true misfit. Large models' value seems limited with poor multilevel data.

Level-specific and aggregate RMSEA likely should not be used with large models, small samples, and skewed data. Single-level RMSEA often is insensitive to misfit with large models (Savalei, 2012). RMSEA always had higher or equal power for small models than large models. Consider RMSEA performance for small Level-1 models with misspecified cross-loadings. Power was always higher for Level-1 RMSEA than aggregate RMSEA. This result likely occurred because Level-1 RMSEA assessed a smaller model (no Level-2 factors, 2 Level-1 factors) than aggregate RMSEA (2 Level-2 factors, 2 Level-1 factors). Level-2 RMSEA's power greatly decreased with skewed data and small samples. Single-level DWLS RMSEA usually has low power with severe non-normality and small samples (Bandalos, 2008).

ICCs impacted Level-2 fit indices differently than Level-1 and aggregate indices. Level-2 fit indices' performance usually improved as ICCs increased (i.e., Level-2 variance increased). Level-1 and aggregate fit indices' performance usually deteriorated as ICCs increased (i.e., Level-1 variance decreased). These findings support previous simulations (Boulton, 2011). As variance decreases, covariances in the covariance matrix decrease in size. Model fit is the alignment between the observed covariance matrix and model-implied covariance matrix. Differences between observed and model-implied covariances indicate misfit. As these differences (i.e., residuals) increase, fit indices usually will indicate worse fit. The smaller the observed covariances, the less potential discrepancy there is between model-implied and observed covariances. As variance decreases, fit indices usually will indicate increasingly good fit. I found that fit indices indicated good fit with low variance (e.g. small ICCs for Level-2 fit) regardless of misfit.

Alternatives to Fit Indices and MSEM

My results generally suggested that Level-2 fit indices could not be trusted when estimating MSEMs using DWLS. These results suggest limitations of DWLS, fit indices, MSEM, and cutoff values. These limitations suggest that alternatives to fit indices and MSEM may be worth considering.

Methodologists have suggested fit index alternatives. Instrumental variable estimators (Bollen, 1996, in press; Bollen, Kirby, Curran, Paxton, & Chen, 2007) have shown increased sensitivity to misfit over popular estimators like ML. These estimators potentially can target sources of misfit and can model categorical or continuous data. Inspecting the expected parameter change in conjunction with modification indices and the modification indices' power to detect misfit has been suggested (Saris, Satorra, & van der Veld, 2009). This approach assesses whether parameter estimates will change meaningfully given model modifications. It emphasizes size of model misspecification and can indicate if statistically non-significant modification indices occur because of insufficient power or an appropriate model.

MSEM alternatives are available. Design-based CFAs are interpreted identically to CFAs, but statistically adjust chi-square and standard errors to reflect nested data. Design-based CFAs are appropriate when only Level-1 inferences are of interest. Stapleton, McNeish, and colleagues provide an overview of these models and technical references (McNeish, Stapleton, & Silverman, 2017; Stapleton, McNeish, & Yang, 2016; Stapleton, Yang, & Hancock, 2016). Yuan and Bentler (2007) suggested fitting separate

CFAs to Level-1 and Level-2 covariance matrices. This approach requires ML because DWLS and instrumental variable estimators cannot analyze summary statistics.

Study Limitations and Future Research

Reliability has received limited evaluation in MSEM simulation studies (but see Geldhof et al., 2014). Several things can influence reliability including the size of factor loadings, ratio of systematic variance to error variance, and the number of indicators. This study did not assess reliability's potential impact. MSEM simulation studies usually hold number of indicators constant. MSEM simulations often vary ICCs by 1) decreasing Level-1 factor variances and/or error variances and increasing Level-2 factor variances and/or error variances or 2) decreasing Level-1 factor loadings and increasing Level-2 factor loadings. In either approach, factor variances and error variances or factor loadings are higher at Level-1 than Level-2. Some MSEM simulations fix factor loadings to 1 across and within levels (e.g. Depaoli & Clifton, 2015). Future research could explore the impact of reliability and these different approaches to adjusting ICCs on parameter estimation and fit indices.

My study suggested Level-2 fit index limitations. These general problems with Level-2 fit indices should be addressed first. If these general problems are addressed, further fit index performance questions can be addressed.

The results provided insight about level-specific fit indices' functioning in relation to their commonly accepted cut-off criteria based on single-level simulations. Future studies can examine the distribution of fit index results across a variety of

conditions to assess if new criteria could be developed. That line of inquiry was not the current study's intent, but it is highly related and should be examined in future research.

This study and most MSEM simulations use equal group sizes across all groups. Unequal group sizes across groups are common in MSEM applications (e.g., Sessoms & Willse, 2019). Previous MSEM simulations on unbalanced groups on parameter estimates and Type I error rates are mixed (Hox & Maas, 2001; Lüdtke et al., 2008). Future research could evaluate the impact of unequal group sizes on level-specific fit indices.

My study fit MCFAs. MCFA is a special case of MSEM, where all factor covariances are estimated freely and no regression hypotheses among common factors are assessed. Some MSEM applications assess mediation or moderation (Preacher et al., 2010, 2016). The level-specific indices I tested simultaneously assess measurement and structural fit, which should be inspected separately. McNeish and Hancock's (2018) structural fit indices could be extended to level-specific indices.

REFERENCES

Asparouhov, T., & Muthén, B. (2007). Computationally efficient estimation of multilevel high-dimensional latent variable models. *Section on Statistics in Epidemiology*, 2531-2535.

Bandalos, D. L. (2008). Is parceling really necessary? A comparison of results from item parceling and categorical variable methodology. *Structural Equation Modeling: A Multidisciplinary Journal*, *15*, 211-240.

Bandalos, D. L. (2014). Relative performance of categorical diagonally weighted least squares and robust maximum likelihood estimation. *Structural Equation Modeling: A Multidisciplinary Journal*, *21*, 102-116.

Beauducel, A., & Herzberg, P. Y. (2006). On the performance of maximum likelihood versus mean and variance adjusted weighted least squares estimation in CFA. *Structural Equation Modeling*, *13*, 186–203.

Bollen, K. A. (1996). An alternative two stage least squares (2SLS) estimator for latent variable equations. *Psychometrika*, *61*, 109-121.

Bollen, K. A. (in press). Model implied instrumental variables (MIIVs): An alternative orientation to structural equation modeling. *Multivariate Behavioral Research*.

Bollen, K. A., Kirby, J. B., Curran, P. J., Paxton, P. M., & Chen, F. (2007). Latent variable models under misspecification: Two-stage least squares (2SLS) and maximum likelihood (ML) estimators. *Sociological Methods & Research, 36,* 84-86.

Boulton, A. (2011). Fit index sensitivity in multilevel structural equation modeling. Unpublished dissertation.

Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (2nd ed.). New York, NY: Guilford Press.

Cheung, M., & Au, K. (2005). Applications of multilevel structural equation modeling to crosscultural research. *Structural Equation Modeling: A Multidisciplinary Journal*, *12*, 598-619. Clifton, J., & Depaoli, S. (2017). Implementing multilevel structural equation models through Bayesian and frequentist estimation: Simulation and applications. Unpublished manuscript.

Depaoli, S., & Clifton, J. P. (2015). A Bayesian approach to multilevel structural equation modeling with continuous and dichotomous outcomes. *Structural Equation Modeling: A Multidisciplinary Journal*, *22*, 327-351.

Depaoli, S., & van de Schoot, R. (2016). Improving transparency and replication in Bayesian statistics: The WAMBS-Checklist. *Psychological Methods*.

Finney, S. J., & DiStefano, C. (2013). Nonnormal and categorical data in structural equation modeling. In G.R. Hancock & R.O. Mueller (Eds.). *A second course in structural equation modeling (2nd ed., pp.* 439-492*).* Charlotte, NC: Information Age.

Finney, S. J., DiStefano, C., & Kopp, J. P. (2016). Overview of estimation methods and preconditions for their application with structural equation modeling. In K. Schweizer and C. DiStefano (Eds.) *Principles and methods of test construction: Standards and recent advances* (pp. 135-165). Boston, MA: Hogrefe.

Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation: Confirmatory factor analysis with ordinal data. *Psychological Methods, 9,* 466-491.

Geldhof, G. J., Preacher, K. J., & Zyphur, M. J. (2014). Reliability estimation in a multilevel confirmatory factor analysis framework. *Psychological Methods, 19,* 72-91.

Guenole, N. (2016). The importance of isomorphism for conclusions about homology: A Bayesian multilevel structural equation modeling approach with ordinal indicators. *Frontiers in Psychology, 7,* 1-17.

Holtmann, J., Koch, T., Lochner, K., & Eid, M. (2016). A comparison of ML, WLSMV, and Bayesian estimation methods for multilevel structural equation models in small samples: A simulation study. *Multivariate Behavioral Research*.

Hox, J. J. (2010). *Multilevel analysis: Techniques and applications* (2nd ed). New York, NY: Routledge.

Hox, J., & Maas, C. (2001). The accuracy of multilevel structural equation modeling with pseudobalanced groups and small samples. *Structural Equation Modeling: A Multidisciplinary Journal, 8,* 157-174.

Hox, J., Maas, C., & Brinkhui, M. (2010). The effect of estimation method and sample size in multilevel structural equation modeling. *Satistica Neerlandica, 64,* 157-210.

Hox, J., van de Schoot, R., & Matthijsse, S. (2012). How few countries will do? Comparative survey analysis from a Bayesian perspective. *Survey Research Methods, 6,* 87-93.

Hsu, H-Y., Kwok, O., Lin, J. H., & Acosta S. (2015). Detecting misspecified multilevel structural equation models with common fit indices: A monte carlo study. *Multivariate Behavioral Research, 50,* 197-215.

Hsu, H-Y, Lin, J-H, Kwok, O-M, Acosta, S., & Willson, V. (2017). The impact of intraclass correlation on the effectiveness of level-specific fit indices in multilevel structural equation modeling; A monte carlo study. *Educational and Psychological Measurement, 77,* 5-31.

Hu, L., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods, 3,* 424-453.

Hu, L., Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6,* 1-55.

Julian, M. W. (2001). The consequences of ignoring multilevel data structures in nonhierarchical covariance modeling. *Structural Equation Modeling: A Multidisciplinary Journal, 8,* 325- 352.

Kim, E. S., Dedrick, R. F., Cao, C., & Ferron, J. M. (2016). Multilevel factor analysis: Reporting guidelines and a review of reporting practices. *Multivariate Behavioral Research, 51,* 881-898.

Kim, E. S., Kwok, O., & Yoon, M. (2012). Testing factorial invariance in multilevel data: A monte carlo study. *Structural Equation Modeling, 19,* 250-267.

Kline, R. B. (2011). *Principles and practice of structural equation modeling* (3rd ed.). New York, NY: Guilford Press.

Lei, P. (2009). Evaluating estimation methods for ordinal data in structural equation modeling. *Quality & Quantity, 43,* 495-507.

Lei, P., & Wu, Q. (2012). Estimation in structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 164-179). New York, NY: Guilford Press.

Li, C.-H. (2016). The performance of ML, DWLS, and ULS estimation with robust corrections in structural equation models with ordinal variables. *Psychological Methods, 21,* 369-387.

Lüdtke, O., Marsh, H. W., Robitzsch, A., Trautwein, U., Asparouhov, T., & Muthén, B. (2008). The multilevel latent covariate model: A new, more reliable approach to group-level effects in contextual studies. *Psychological Methods, 13,* 203-229.

Lüdtke, O., Marsh, H. W., Robitzsch, A., & Trautwein, U. (2011). A 2x2 taxonomy of multilevel latent contextual models: Accuracy-bias tradeoffs in full and partial error correction models. *Psychological Methods, 16,* 444-467.

McNeish, D. & Hancock, G. R., (2018). The effect of measurement quality on targeted structural model fit indices: A comment on Lance, Beck, Fan, and Carter (2016). *Psychological Methods*, *23*, 184-190.

McNeish, D., Stapleton, L. M., & Silverman, R. D. (2017). On the unnecessary ubiquity of hierarchical linear modeling. *Psychological Methods*, *22*, 114-140.

Muthén, B. O. (1991). Multilevel factor analysis of class and student achievement components. *Journal of Educational Measurement*, *28*, 338-354.

Muthén, B. O. (1994). Multilevel covariance structure analysis. *Sociological Methods & Research*, *22*, 376-398.

Meuleman, B., & Billiet, J. (2009). A Monte Carlo sample size study: How many countries are needed for accurate multilevel SEM? *Survey Research Methods*, *3*, 45-58.

Muthén, B. O., & Asparouhov, T. (2008). Growth mixture modeling: Analysis with non-Gaussian random effects. In G. Fitzmaurice, M. Davidian, G. Verbeke, & G. Molenberghs (Eds.), *Longitudinal data analysis* (pp. 143-165). Boca Raton, FL: Chapman & Hall/CRC.

Muthén, B. O., du Toit, S. H., & Spisic, D. (1997). Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes. Unpublished document. Retrieved from http://statmodel.com/SEM.shtml/

Muthén, L. K., & Muthén, B. O. (1998-2017). Mplus user's guide (8th ed.). Los Angeles, CA: Muthén & Muthén.

Oranje, A. (2003, April). *Comparison of estimation methods in factor analysis with categorized variables: Applications to NAEP data*. Paper presented at the annual conference of the American Educational Research Association, Chicago, Illinois.

Peugh, J. L. (2010). A practical guide to multilevel modeling. *Journal of School Psychology*, *48*, 85-112.

Preacher, K. J., Zhang, Z., & Zyphur, M. J. (2011). Alternative methods for assessing mediation in multilevel data: The advantages of multilevel SEM. *Structural Equation Modeling: A Multidisciplinary Journal*, *18*, 161-182.

Preacher, K. J., Zhang, Z., & Zyphur, M. J. (2016). Multilevel structural equation models for assessing moderation within and across levels of analysis. *Psychological Methods*, *21*, 189-205.

Preacher, K. J., Zyphur, M. J., & Zhang, Z. (2010). A general multilevel SEM framework for assessing multilevel mediation. *Psychological Methods*, *15*, 209-233.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage Publications.

Ryu, E., & West, S. G. (2009). Level-specific evaluation of model fit in multilevel structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *16*, 583- 601.

Saris, W. E., Satorra, A., & van der Veld, W. M. (2009). Testing structural equation models or detection of misspecifications? *Structural Equation Modeling*, *16*, 561-582.

Savalei, V. (2014). Understanding robust corrections in structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *21*, 149-160.

Sessoms, J., Finney, S. J., & Kopp, J. P. (2016). Does the measurement or magnitude of academic entitlement change over time? *Measurement and Evaluation in Counseling and Development*, *43*, 243-257.

Sessoms, J., & Willse, J. T. (2019). *A tutorial on multilevel confirmatory factor analysis*. Revised manuscript submitted for publication.

Stapleton, L. M. (2013). Multilevel structural equation modeling with complex sample data. In G. R. Hancock and R. M. Mueller (Eds.) *Structural equation modeling: A second course* (2nd ed.) (pp. 51-562.). Charlotte, NC: Information Age Publishing.

Stapleton, L. M., McNeish, D. M., & Yang, J. S. (2016). Multilevel and single-level models for measured and latent variables when data are clustered. *Educational Psychologist, 51*, 317-330.

Stapleton, L. M., Yang, J. S., & Hancock, G. R. (2016). Construct meaning in multilevel settings. *Journal of Educational and Behavioral Statistics*, *41*, 481-520.

Yang-Wallentin, F., Jöreskog, K. G., & Luo, H. (2010). Confirmatory factor analysis of ordinal variables with misspecified models. *Structural Equation Modeling: A Multidisciplinary Journal, 17*, 392-423.

Yuan, K.-H., & Bentler, P. M. (2007). Multilevel covariance structure analysis by fitting multiple single-level models. *Sociological Methodology*, *37*, 53-82.

Yu, C., & Muthén, B. (2002, April). *Evaluation of model fit indices for latent variable models with categorical and continuous outcomes*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Zyphur, M. J., Kaplan, S. A., & Christian, M. S. (2008). Assumptions of cross-level measurement and structural invariance in the analysis of multilevel data: Problems and solutions. *Group Dynamics: Theory, Research, and Practice, 12*, 127-140.

APPENDIX A

LARGE MODEL CONVERGENCE FOR SKEWED DATA

Fit Model	L1 Misp. Cross-loading	L1 Misp Factor	L2 Misp Cross-loading	L2 Misp Factor	L1 & L2 Misp Cross-loading	L1 & L2 Misp Factor	True Mdl	Data
L1 ind L2 sat	1	1	1	1	1	1	1	50groups icc15
L1 theo L2 sat	1	1	1	1	1	1	1	50groups icc15
L1 sat L2 theo	1	1	1	1	1	1	1	50groups icc15
L1 sat L2 theo	.94	.94	1	.99	1	.99	.94	50groups icc15
L1 theo L2 theo	.93	.93	1	.99	1	.99	.93	50groups icc15
L1 ind L2 sat	1	1	1	1	1	1	1	100 group icc15
L1 theo L2 sat	1	1	.99	1	1	.99	1	100group icc15
L1 sat L2 theo	1	1	1	1	1	1	1	100group icc15
L1 sat L2 theo	.98	.98	1	1	1	1	.98	100group icc15
L1 theo L2 theo	.99	.99	1	1	1	1	.99	100group icc15
L1 ind L2 sat	1	1	1	1	1	1	1	50groups icc30
L1 theo L2 sat	1	1	1	1	1	1	1	50groups icc30
L1 sat L2 theo	1	1	1	1	1	1	1	50groups icc30
L1 sat L2 theo	.99	.99	1	1	1	1	.99	50groups icc30
L1 theo L2 theo	.99	.99	1	1	1	1	.99	50groups icc30
L1 ind L2 satu	1	1	1	1	1	1	1	100group icc30

L1 theo L2 sat	1	1	1	1	1	1	1	100group icc30
L1 sat L2 theo	1	1	1	1	1	1	1	100group icc30
L1 sat L2 theo	1	1	1	1	1	1	.99	100group icc30
L1 theo L2 theo	.99	1	.99	1	1	1	1	100group icc30

APPENDIX B

LARGE MODEL CONVERGENCE FOR SYMMETRIC DATA

Fit Model	L1 Misp. Cross-load	L1 Misp Factor	L2 Misp Cross-load	L2 Misp Factor	L1 & L2 Misp Cross-load	L1 & L2 Misp Factor	True Mdl	Data
L1 ind L2 sat	1	1	1	1	1	1	1	50groups icc15
L1 theo L2 sat	.99	1	1	1	1	1	1	50groups icc15
L1 sat L2 theo	1	1	1	1	1	1	1	50groups icc15
L1 sat L2 theo	.95	.95	1	1	1	1	.95	50groups icc15
L1 theo L2 theo	.96	.95	1	1	1	.99	.95	50groups icc15
L1 ind L2 sat	1	1	1	1	1	1	1	100 group icc15
L1 theo L2 sat	1	1	1	1	1	1	1	100group icc15
L1 sat L2 theo	1	1	1	1	1	1	1	100group icc15
L1 sat L2 theo	1	1	1	1	1	1	1	100groups icc15
L1 theo L2 theo	1	1	1	1	1	1	.99	100group icc15
L1 ind L2 sat	1	1	1	1	1	1	1	50groups icc30
L1 theo L2 sat	1	1	1	1	1	1	1	50groups icc30
L1 sat L2 theo	1	1	1	1	1	1	1	50groups icc30
L1 sat L2 theo	.99	.99	1	1	1	1	.99	50groups icc30
L1 theo L2 theo	.99	.99	1	1	1	1	.99	50groups icc30
L1 ind L2 satu	1	1	1	1	1	1	1	100group icc30

L1 theo L2 sat	1	1	1	1	1	1	1	100group icc30
L1 sat L2 theo	1	1	1	1	1	1	1	100group icc30
L1 sat L2 theo	1	1	1	1	1	1	1	100group icc30
L1 theo L2 theo	1	1	1	1	1	1	1	100group icc30

APPENDIX C

SMALL MODEL CONVERGENCE FOR SKEWED DATA

Fit Model	L1 Misp. Cross-load	L1 Misp Factor	L2 Misp Cross-load	L2 Misp Factor	L1 & L2 Misp Cross-load	L1 & L2 Misp Factor	True Mdl	Data
L1 ind L2 sat	1	1	1	1	1	1	1	50groups icc15
L1 theo L2 sat	1	1	1	1	1	1	1	50groups icc15
L1 sat L2 theo	1	1	1	1	1	1	1	50groups icc15
L1 sat L2 theo	.95	.95	1	1	1	1	.95	50groups icc15
L1 theo L2 theo	.95	.95	1	1	1	1	.95	50groups icc15
L1 ind L2 sat	1	1	1	1	1	1	1	100 group icc15
L1 theo L2 sat	1	1	1	1	1	1	1	100group icc15
L1 sat L2 theo	1	1	1	1	1	1	1	100group icc15
L1 sat L2 theo	1	1	1	1	1	1	1	100group s icc15
L1 theo L2 theo	1	1	1	1	1	1	1	100group icc15
L1 ind L2 sat	1	1	1	1	1	1	1	50groups icc30
L1 theo L2 sat	1	1	1	1	1	1	1	50groups icc30
L1 sat L2 theo	1	1	1	1	1	1	1	50groups icc30
L1 sat L2 theo	1	1	1	1	1	1	1	50groups icc30
L1 theo L2 theo	1	1	1	1	1	1	1	50groups icc30
L1 ind L2 satu	1	1	1	1	1	1	1	100group icc30

L1 theo L2 sat	1	1	1	1	1	1	1	100group icc30
L1 sat L2 theo	1	1	1	1	1	1	1	100group icc30
L1 sat L2 theo	1	1	1	1	1	1	1	100group icc30
L1 theo L2 theo	1	1	1	1	1	1	1	100group icc30

APPENDIX D

SMALL MODEL CONVERGENCE FOR SYMMETRIC DATA

Fit Model	L1 Misp. Cross-load	L1 Misp Factor	L2 Misp Cross-load	L2 Misp Factor	L1 & L2 Misp Cross-load	L1 & L2 Misp Factor	True Mdl	Data
L1 ind L2 sat	1	1	1	1	1	1	1	50groups icc15
L1 theo L2 sat	1	1	1	1	1	1	1	50groups icc15
L1 sat L2 theo	1	1	1	1	1	1	1	50groups icc15
L1 sat L2 theo	.98	.98	1	1	1	1	.98	50groups icc15
L1 theo L2 theo	.98	.98	1	1	1	1	.98	50groups icc15
L1 ind L2 sat	1	1	1	1	1	1	1	100 group icc15
L1 theo L2 sat	1	1	1	1	1	1	1	100group icc15
L1 sat L2 theo	1	1	1	1	1	1	1	100group icc15
L1 sat L2 theo	1	1	1	1	1	1	1	100group s icc15
L1 theo L2 theo	1	1	1	1	1	.99	.99	100group icc15
L1 ind L2 sat	1	1	1	1	1	1	1	50groups icc30
L1 theo L2 sat	1	1	1	1	1	1	1	50groups icc30
L1 sat L2 theo	1	1	1	1	1	1	1	50groups icc30
L1 sat L2 theo	1	1	1	1	1	1	1	50groups icc30
L1 theo L2 theo	1	1	1	1	1	1	1	50groups icc30
L1 ind L2 satu	1	1	1	1	1	1	1	100group icc30

L1 theo L2 sat	1	1	1	1	1	1	.99	100group icc30
L1 sat L2 theo	1	1	1	1	1	1	1	100group icc30
L1 sat L2 theo	1	1	1	1	1	1	1	100group icc30
L1 theo L2 theo	1	1	1	1	1	1	1	100group icc30

APPENDIX E

EMPIRICAL ASYMMETRY CATEGORY PROPORTIONS

Item	Response Category	Proportion of Responses
1	a	.098
	b	.115
	c	.108
	d	.679
2	a	.094
	b	.105
	c	.099
	d	.702
3	a	.105
	b	.107
	c	.107
	d	.682
4	a	.105
	b	.106
	c	.104
	d	.685
5	a	.102
	b	.094
	c	.114
	d	.690
6	a	.110
	b	.102
	c	.105
	d	.684
7	a	.98
	b	.112
	c	.102
	d	.688
8	a	.116
	b	.110
	c	.101
	d	.673

Note. These responses come from one replication of the small model, asymmetric data, many groups, and large ICCs condition. The intended category proportions based on the data generating model were .1, .1, .1, and .7 (for categories A, B, C, and D, respectively).